自然言語処理 41－6
（1984. 1. 26）

確率的言語処理へのアプローチ

藤崎哲之助
日本アイ・ビー・エム(株) サイエンス・インスティテュート

1. はじめに

自然言語の解析を行うために文法を作る試みは数多く存在するが、対象とする文に作為的な制限を行わない場合、自然言語のあいまい土が常に問題となる。Church, Ramesh [1] らの報告によれば、彼らの英文解析例でのあいまい度は最大958 ("Inasmuch as allocating cost is a tough job I would like to have the total costs related to each product.") に達し、約2%の文が300以上のあいまいな解釈を有したと報告されている。

これは特に驚くべき事ではない。すなわち、英語の場合を例にとれば、前置詞句の後端、接続詞句の両端などの名詞句の境界は常にあいまい土を生じさせる可能性がある。従って、文のあいまい度を、そのようなあいまい土を生じさせる構造の数の関数として推定することができる。

CATLAN数 [2] は、N項式にカッコを挿入する組み合わせの数で指数的に増大するが、C(1)=1, C(2)=2, C(3)=5, 14, 42, 132, 469, 1430, 4892, ・・・ 自然言語のあいまい度をよく近似する。例えば、次の例文：

I saw a man in a park with a scope.

は5つのあいまいな解釈を持つが、これは片側あいまいの3つの句構造, saw NP, in NP, with NP, を持つためで、C(3)=5と一致する。

このような自然言語のあいまい土を統語的立場だけから完全に対処するのは不可能で、意味上、運用上での制約を導入するのが長い人工知能の歴史の教えである。すなわち、意味ネットワーク、意味フレームなどにより、対象世界を記述し、それらの上での意味的、運用的、あるいは常識的制約により、その場面での正しい解釈を決定することが行われる。

このような立場は心理学、認知科学の立場からも認められつつあり、終局的にはこのような方向で人間の言語理解活動がモデル化されると筆者は考えているが、現実的には、意味的・運用的知識を始めとして、常識、果ては話者の好み、性格に至るまで形式的にデータ・ベースにネットワークとして記述することの困難には、通り知れないものがある。そして、現在に至るまでも、ごく限られた実用的に至らない範囲で成功しているにすぎない。[3,4]

さらに、データ・ベースの照会の場面での次の例文を考える。

print for me the sales of stair carpets.

この文は一見あいまいでなく見えるが、データ・ベース中に米国の州別売り上げが情報として含まれるならば、"ME" としてMaine州の解釈をすることができ、文の真意はあいまいとなる。[5] この例の場合、真意は質問者のくせ、傾向を知った上で本質的に確率的にしか決定し得ないものと考えられる。

このような2つの立場 (a)現段階の技術として、言語の解析に必要な知識全てを詳細に至るまで記述し上げることには困難がある。(b)言語の解析に本質的に確率的判断も必要とされる場面もある。) から、筆者は、確率的文法およびそれを用いた確率的構文解析[6]が有効であると考える。本稿では、

文法を自動的に確率化する手法の概略、またその結果得られた確率的文法による確率的構文解析の有効性を、英語の構文解析の例により示す。

2. 確率的構文解析

文脈自由文法の構文解析では、各解析木は構文規則適用の連鎖として表現できる。従って、あらかじめ各構文規則に確率が与えられていれば、(各確率は左辺の等しいもので正規化されているものとする。[6])それらの積として、解析木の出現確率を得ることができる。この確率は、その文がその解析木に対応した真意で発生する同時確率であり、一般には非常に小さい。

文があいまいであるとは解析木が複数個得られることであるので、それらの解析木間の相対確率が、あいまいな解釈における信頼度を与えると考えるのは自然である。従って、各構文規則の確率の全体として、意味的、運用的、あるいは常識、発話者の好み、性格など反映したものとして設定できれば、確率的構文解析が有効なあいまい土除去の手段となる。

このような枠組みであいまい土の除去を行うには、各規則に与えられる確率として種々の要因を考慮した精確なものを与える必要があり、各確率を経験的に与えたり、ad hocなものとして断片的に与えることはできない。本稿の次節以降では、それらの規則毎の確率を訓練用の例文の集合より自動的に推定する方式を紹介する。この方式によれば、文法の確率化が正確に行なえるばかりでなく、訓練用の文の選出域をうまく設定することにより、任意の対象分野で最適な文法の確率化を行うことができる。発話者のくせ、好みなども訓練域の設定により吸収することができる。図1に例文よりの文法の確率化、その確率化された文法を用いての確率的構文解析の関連を示す。

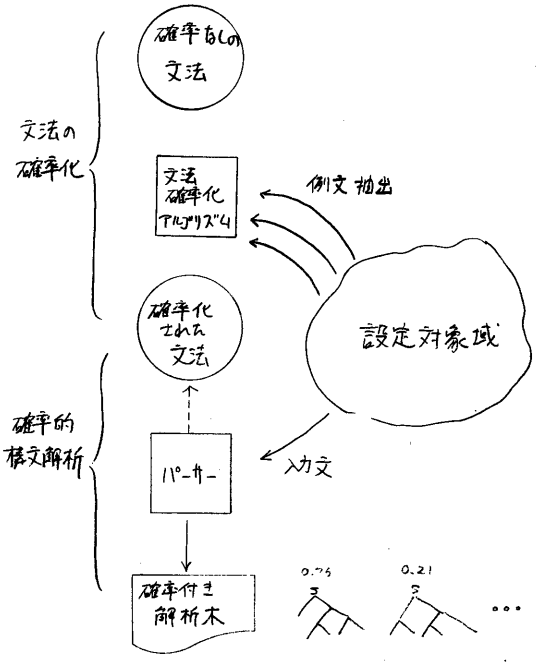

図1

3. マルコフ・モデルのパラメータ自動推定

本稿で扱うのは文脈自由文法の確率化であるが、準備として、マルコフ・モデルのパラメータ(遷移確率)の自動推定手続の概略を紹介する。

マルコフ・モデルでは、状態S間に遷移πが定義されており、遷移の際に文字b_{st}が出力される。そして、各遷移に対して遷移確率$g(s,t)$:状態Sで遷移πの起る確率($\sum_t g(s,t)=1$)が定義される。

このようなマルコフ・モデルにおける遷移確率$\{g(s,t)\}$は、出力文字列の集合$\{B_1, B_2, \cdots\}$から次の手順で推定することができる。

1. 各遷移確率$\{g(s,t)\}$に適当な推定値を与える。
2. そのマルコフ・モデルを線形文法に見

立て、各出力文字列 D_i を構文解析する。得られる解析木はマルコフ・モデル上での状態の遷移列であるが、それを D_{ij}（j はあいまいな解析木を識別する標識とする。）とする。

3. 各遷移列 D_{ij} の出現確率は、その D_{ij} に含まれる遷移に対応づけられた $g(s,t)$ の積として計算できるのでそれを計算し、$Pr(i,j)$ とする。

4. この $Pr(i,j)$ を用いて、D_i を発生する際に状態 S からの遷移 α をマルコフ・モデル上で通過した回数 $C(s,t,i)$ を推定することができる。

$$C(s,t,i) = \frac{\sum_j Pr(i,j) * n(s,t,D(i,j))}{\sum_j Pr(i,j)}$$

但し $n(s,t,D(i,j))$ は D_{ij} が $\langle s,t \rangle$ の遷移を含む度数

5. $C(s,t,i)$ を正規化し、i で平均することにより、$g(s,t)$ の新しい推定値を得ることができる。

$$new\ g(s,t) = \frac{\sum_i C(s,t,i)}{\sum_{u,i} c(s,u,i)}$$

6. $g(s,t)$ を $new\ g(s,t)$ で置き換え、収束していなければ2へ戻る。

以上の手続は収束することが証明されており[7]、例文の集合から遷移確率を得ることができる。本稿では、この手続を直観的にわかりやすい形で紹介したが、より計算量の少なくてすむ方式が[8,9]に紹介されている。

4. 文脈自由文法の確率化への拡張

前節で述べたマルコフ・モデルの遷移確率の推定手法を文脈自由文法の確率化へ拡張することができる。

与えられた文脈自由文法 G に対する可能な文型式（Sentential Form）を状態とし、文脈自由の構文規則の適用を状態の遷移とするようなマルコフ・モデルを考える。（図2）

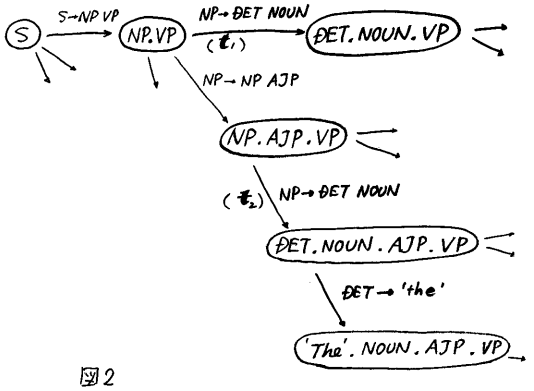

図2

このモデルでは、例えば状態 'S' が規則 'S→NP VP' の適用により状態 'NP.VP' に遷移し、状態 'NP.VP' は規則 'NP→DET NOUN' の適用により状態 'DET.NOUN.VP' に遷移する。

このようなマルコフ・モデルを想定すれば、文法 G から生成された文を集めて、3節の手続きより、各遷移確率の推定を行うことができる。ただし、文脈自由文法では確率は各構文規則毎に与えられるので確率は規則毎に計算されねばならない。（図2での α_1 と α_2 の遷移確率は同じにならねばならない。）そのため前節の手順4で得られる $\langle s,t \rangle$ の遷移の期待通過回数 $C(s,t,i)$ は各構文規則毎にまとめられ、また手順5の正規化も、構文規則上での左辺の等しいものの間で行われることになる。（この変更をはずせば文脈依存の文法にも適用の可能性はある。）

5. 実験

実験のベースとなる文法として、1960年代にハーバード大学の久野氏により作成された文法[10,11]を用いた。この文法は拡張されたGraibach標準形[12]の規則2118より構成されるが、後にのべる構文解析プログラムの都合より5241

のChomsky標準形の規則に変換した。
特にこの文法をベースとして選んだのは次の3つによる：

a. 広範囲の英文を解析する能力があることが報告されている。
b. 純粋な形の文脈自由文法に変換可能であり、我々の方式に好都合であった。
c. 構文的、意味的差異により各範ちゅうが細分されており、解釈の意味的差異が解析木の構造的差異に反映する。(形容詞7分類、副詞8分類、BE-動詞3分類、自動詞3分類，他動詞7分類など)

辞書としては、久野文法に付随した約2万5千語の辞書と、ウェブスター第7版(約12万語)の辞書を用いた。
構文解析プログラムとしては、この実験のためにCocke-Kasami-Young型のパーサー[123]を作成した。あいまい度の高い文を効率に解析する必要から、効率の向上に考慮を払った。
最終的にこの構文解析プログラムは、DATAMATION, READER'S DIGEST 合わせて31の記事から選んだ文を表1の性能で解析することができた。

総文数	5528
平均語長	10.85
解析できた文の数	3582
1文当りの処理時間 (3033 Uni Processor,) CPU時間	0.89秒
平均あいまい度	48.5

特にこのプログラムでは高速化を実現するため、ハッシングを効用し、また以下に述べる部分木の共有化を可能な限り実現する工夫を行った。(構造木共有型パーザー)
すなわち構文木の表現に従来用いられるn本足のノード(親子関係を表現する。)に加えてOR-リンク(ある非末端記号の下記構造に異なる部分構文木があいまいに存在するとき、それらの部分構文木をリストとして表現する。)を導入した。例えば、図3(a)のNPの例に見られるように'ABC'の部分に2つの異なるNPが作られるとき、それらの下位部分木をOR-リンクでつないで図3-(b)の構造の表現が行われる。

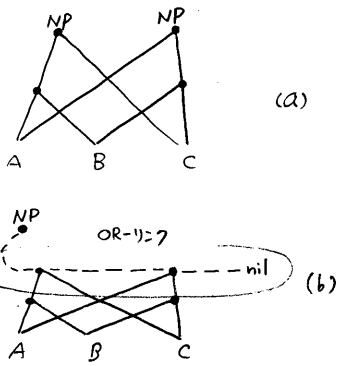

図3

このOR-リンクの導入により、規則適用の対象となる中間ノードの数が減少し、CKY型のアルゴリズムのパーサーでは$O(N^2)$の高速化が期待できる。(但しNはOR-リンクにつながれる部分木の平均個数)従って、あいまい度の高い文の解析において特に有効である。

確率化。最初に与えられた5241の構文規則、また与えられた5528の文中に観測された辞書規則に対して、確率化を行った。その結果得られた確率的文脈自由文法の一部を図4に示す。図中の数字が反復により各規則に最終的に与えられた確率で、例えば図中(a)，(b)の行は、範ちゅう'IT4'から語彙'HELP'，'SEE'が発生する辞書規則とその確率、(c)，(d)は'SE→PRN VX PD'、'SE→AAA 4X VX PD'の構文規則の確率

```
** IT4
        0.98788          HELP      ---(a)
        0.00931          SEE       ---(b)
        0.00141          HEAR
        0.00139          WATCH
        0.00000          HAVE
        0.00000          FEEL
                  .
                  .
                  .
** IT5
        0.78550          GET
        0.13246          HAVE
        0.04307          WANT
        0.03181          SEE
                  .
                  .
                  .
** SE
        0.28754          PRN VX PD          ---(c)
        0.25530          AAA 4X VX PD       ---(d)
        0.14856          NNN VX PD
        0.13567          AV1 SE
        0.04006          PRE NQ SE
        0.02693          AV4 IX MX PD
        0.01714          NUM 4X VX PD
        0.01319          IT1 N2 PD
                  .
                  .
                  .
** VX
        0.16295          VT1 N2
        0.14372          VI1
        0.11963          AUX BV
        0.10174          PRE NQ VX
        0.09460          BE3 PA
                  .
                  .
                  .
```

図4

がそれぞれ 28.8%, 25.5% であることを示している。 但し：

SE — sentence.
PRN — pronoun.
VX — predicate.
PD — period preceded by optional post sentential modifier.
AAA — article, adjective, etc.
4X — subject noun phrase

図4に示すような確率文法が得られれば、同じ構文解析プログラムを用いて、確率的構文解析を行うことができる。特に、あいまいな解析の起る際、最も発生確率の高いものだけに興味がある場合には、Viterbi [13] の手法を構文解析に適用して、構文解析の終了と共に最大の発生確率を与える構文木を絶対確率、相対確率と共に $O(l^3)$ の手間で得ることができる。（l は入力文の語長）それは図3のOR-リンク上に常に最大確率を与える部分木だけを残すことにより実現でき、その場合、オ2位以降の部分解析木を捨て、そのための記憶域を開放できるため、作業域の減少も実現できる。

図5は、この確率文法を用いて確率的構文解析を行った例である。 入力文に対し、部分木A, B, Cの差異により示される3つのあいまいな解釈が得られている。（A, B, CはOR-リンクにより結ばれている。） この内、Cは通常の解釈に対応し、Aは 'OUTSIDE' と 'ART' の間の関係代名詞 which が省略された構造に対応する。 Bは、'ART SERVICE' が動詞 'UTILIZE' の目的語となり、'OUTSIDE' が前置詞として間に挿入された構造に対応する。

各A, B, Cにつけられた数字、0.356, 0.003, 0.641 は最終的に得られた構文規則の確率より計算した各解析木の相対確率である。 この例に示されるように、一番もっともらしい解釈の順（この場合 C, A, B）に高い値が与えられている。 また カッコ内の数字は反復の途中で得られた規則の確率をもとに計算した解析木の相対確率で、この場合に示されるように、正しい方向に収束していることがわかる。

表2に、結果をより系統的にまとめて示す。 表2の場合1, 場合2は、IBM内部レター、Datamation, Reader's Digest の雑誌をそれぞれ対象域とした実験に対応している。 それぞれの場合で文法規則は確率化され、その後、eに示される数の文を図5に示されるような確率付き構文解析木として出力した。その出力リストで、最も自然な解釈に最高確率が与えられているかを人手で

```
<WE DO NOT UTILIZE OUTSIDE ART SERVICES DIRECTLY . >
** total ambiguity is :      3

   *:              SENTENCE
     *:              PRNOUN        'WE              '
     *:              PREDICATE
       *:              AUXILIARY     'DO              '
       *:              INFINITE VERB PHRASE
         *:              ADVERV TYPE1 'NOT             '
         A: 0.356(0.52)  INFINITE VERB PHRASE
          |*:              VERB TYPE IT1'UTILIZE        '
          |*:              OBJECT
          |  *:              NOUN          'OUTSIDE     '
          |  *:              ADJ CLAUSE
          |    *:              NOUN          'ART       '
          |    *:              PRED. WITH NO OBJECT
          |      *:              VERB TYPE VT1'SERVICES  '
         B: 0.003(0.28)  INFINITE VERB PHRASE
          |*:              VERB TYPE IT1'UTILIZE        '
          |*:              OBJECT
          |  *:              PREPOSITIN    'OUTSIDE     '
          |  *:              NOUN OBJECT
          |    *:              NOUN          'ART       '
          |  *:              OBJECT
          |    *:              NOUN          'SERVICES  '
         C: 0.641(0.20)  INFINITE VERB PHRASE
          |*:              VERB TYPE IT1'UTILIZE        '
          |*:              OBJECT
          |  *:              NOUN          'OUTSIDE     '
          |  *:              OBJECT MASTER
          |    *:              NOUN          'ART       '
          |    *:              OBJECT MASTER
          |      *:              NOUN          'SERVICES  '
     *:              PERIOD
       *:              ADVERB TYPE1 'DIRECTLY       '
       *:              PRD          '.              '
```

図5

確認した。(表2-g)

a.	テストケース	1	2
b.	対象分野	IBM社内レター	雑誌
c.	訓練用文数	624	3582
d.	1文当り平均語長	12.65	10.85
e.	検査された文数	21	63
f.	正しい解析木が含まれなかった文の数	2	4
g.	最高確率が最も自然な文に与えられた	18	54
h.	他の自然でない解釈に最高確率が与えられた	1	5
i	成功率 (g/(g+h))	0.947	0.915

表2

6. 考察

表2に示すように、本方式が、構文解析のあいまいさの除去に十分に有効であることが示せた。

Eは主として辞書の誤りによるものであるが、Fはベースとした文法の次のような構造的問題による場合が半数以上を占めた。

すなわち、英文における TO-不定詞句、前置詞句などのかかり受けは、多くの場合あいまいで、前述のCATLAN数のあいまい度を発生する。久野文法では、それを回避するため、主動詞句以降に限られる副詞や、それらの句は、主文の最後に文の後修飾句(Post-modifying)として、文の終端記号から派生するという形式で処理している。従って、CATLAN数のあいまいさはそれらから生じないが、逆に、それらの被修飾ー修飾の関係は木構造として現われず、それらを解析する手段は失なわれている。例えば、図5の解析木において、副詞の'DIRECTLY'が本来なら'INFINITE VERB PHRASE'の部分木として派生すべきであるが、実際には、'PERIOD'から派生していることがわかる。

ケース1における唯一の誤りもこれに由来している。すなわち、その例では'... is going to work ...'が含まれ、'TO WORK'の部分の解釈として ① TO-不定詞句 ② 前置詞句

の2つの解釈が図6の斜線部で競合している。

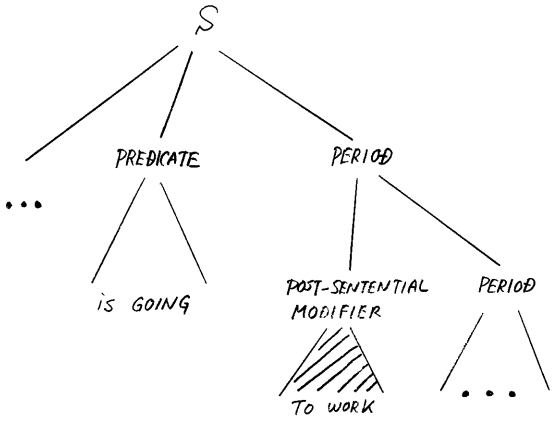

図6

しかし、図6のように'TO WORK'の部分が直前の文脈から分離して最上位から直接派生しているので、実際に競合しているのは、図7の2つの構文規則である。

Post-Sentential Modifier → <TO-不定詞句>

Post-Sentential Modifier → <前置詞句>

図7

前後の文脈を考慮するなら、一般には前置詞句の方が、TO-不定詞句より高頻度で現われ、'TO'は前置詞として頻度が高く、'WORK'も名詞になり得るので図7では2番目の規則に高い確率が与えられるのは自然である。

7. 言語認識型問題への適用の展望

筆者は[14]において漢字仮名混り文の仮名ふりとわかち書きが、Viterbiアルゴリズム[13]で形式化できることを示した。(その後京大の黒橋助教授も追試を行い、その効果を確率して下さった。) これは、言語をマルコフ・モデルを用いてモデル化し、局所的あいまいさをViterbiアルゴリズムで解消する枠組みであり、この漢字仮名変換の他にも、OCRの文字認識、音声認識、ステノ・タイプライター出力の解析の後処理など数多くの応用が行われ、その効果が報告されている。[8,9,15] そして、今後さらに多くの問題への応用が期待される強力な枠組みである。

この枠組みで本質的に必要なものは、与えられた言語での文の発生確率を近似することであり、いままでShannonにより提案された N-gram近似[16]が唯一の手法として用いられてきた。しかしこのN-gram近似は、言語の文脈依存性が局所的である場合には非常に有効であるが、文脈依存の距離をのばすためには、距離に対し指数的な数のマルコフ・モデルの状態を用意する必要があり、より長い文脈依存性の影響を考慮する問題、例えば、仮名漢字変換など、にこのままで適用するには無理がある。一方自然言語の文脈依存性が長い距離にわたる事がよく知られている。(図7)

Is the letter written in English John's?

図7

そこでN-gramによる近似に換わるものとして、本稿で紹介した文脈自由以上の確率文法による文の発生確率の近似が有効であると期待される。今後のこの方面の研究が期待される。

8. あとがき

本研究の主要部分は、筆者がIBMのワトソン研究所に滞在中のプロジェクトとしてJohn Cocke (IBM Fellow)の財政的支援を得て行ったものである。また、ハーバード大学の久野教授には多大の技術

的・精神的支援を得た。また、ワトソン研究所の連続音声認識グループ(F. Jelinek, 他)より多くの技術的コメント・助言を得た。F. Damerau, S. Petrick などのTQAグループより協力を得た。英文法上の問題に関しては、ニューヨーク州大のE. Black氏、実体化に関してはハーバード大のB. Green氏、J. Luty氏の協力を得た。
以上の人々に多大の感謝をする。

参考文献

[1] Church, K., Ramesh, P., "Coping with Syntactic Ambiguitiy", MIT, Laboratory for Computer Science Memo, April, 1982
[2] Knuth, D., Fundamental Algorithms, Vol 1. in The Art of Computer Programming, Addison Wesley, 1975
[2] Martin, W., et. al., Preliminary Analysis of a Breadth-First Parsing Algorithm: Theoretical and Experimental Results, MIT LCS report TR-261, MIT 1981
[3] Winograd, T., Understanding Natural Language, Academic Press, 1972
[4] Woods, W., The Lunar Sciences Natural Language Information System, BBN Report No. 2378, Bolt, Beranek and Newman
[5] Martin, W., et. al., Preliminary Analysis of a Breadth-First Parsing Algorithm: Theoretical and Experimental Results, MIT LCS report TR-261, MIT 1981
[6] Fu, K. S., Syntactic Methods in Pattern Recognition, Vol 112, Mathematics in science and Engineering, Academic Press, 1974
[7] Baum, L. E., A Maximazation Technique occurring in the Statistical Analysis of Probablistic Functions of Markov Chains, Vol, 41, No. 1, The Annals of Mathematical Statistics, 1970
[8] Bahl, L., Jelinek, F., and Mercer, R., A Maximum Likelihood Approarch to Continuous Speech Recognition, Vol. PAMI-5, No. 2, IEEE Trans. Pattern Analysis and Machine Intelligence, 1983
[9] Jelinek, F., "Continuous Speech Recognition by Statistical Method", Proc. of IEEE, Vol 64, No. 4, 1976
[10] Kuno, S., The Augmented Predictive Analyzer for Context-free Languages-Its Relative Efficiency, Vol. 9, No. 11, CACM, 1966
[12] Kuno, S., Oettinger, A. G., Syntactic Structure and Ambiguity of English, Proc. FJCC, AFIPS, 1963
[13] Forney, G. D. Jr., "The Viterbi Algorithm", Proc. of IEEE, Vol. 61, 1973
[14] 藤崎, "動的計画法による漢字仮名混り文の単位切りと仮名ふり", 情報処理学会, 自然言語研究会 28-5, 1981
[15] Raviv, J., "Decision Making in Markov Chains Applied to the Problem of Pattern Recognition, IEEE Trans. Information Theory, Vol IT-3, No. 4, 1967
[16] Shannon, C. E., "Prediction and Entropy of printed English", Bell Sys. J, vol. 30, 1951