ゲーム情報学 2−11
(2000. 3. 17)

棋譜テキストからの特徴素抽出とそれを用いた棋譜の分類

中村 貞吾, 梶山 貴司

九州工業大学 情報工学部 知能情報工学科
E-mail: {teigo,kajiyama}@dumbo.ai.kyutech.ac.jp

概要

囲碁棋譜を着手毎の着点が符号化されてできた文字列（棋譜テキスト）であるととらえ，その部分文字列を特徴素とするベクトル空間法によって棋譜の内容に関する分類を行なう．

日本棋院「棋譜データ集 96」CD-ROM に収録されている 13,232 棋譜（学習用棋譜 10,044, テスト用棋譜 2,188）に対して,「盤上の位置」,「直前の着手との位置関係」,「着点の周囲の状況」の３つの観点に基づく６通りの着手符号化法と n-gram 文字列の出現頻度情報を用いた４通りの特徴素選出法の組に対して棋譜分類実験を行ない，どのような特徴が棋譜の分類に有効であるかを検証した．

Feature Extraction from Encoded Texts of Moves and Categorization of Game Records

Teigo NAKAMURA, Takashi KAJIYAMA

Department of Artificial Intelligence, Kyushu Institute of Technology
E-mail: {teigo,kajiyama}@dumbo.ai.kyutech.ac.jp

Abstract

Game records can be regarded as text strings by suitable encoding from each move to a character. In this paper, we investigate what information should be included in one code and how to extract index terms from the texts in order to analyze and classify these texts of game records appropriately. We show some experimental results of classifying game records in "Kifu Database 96".

1 はじめに

一局のゲームは，個々の着手の着手位置を記録した棋譜によって記述され，棋譜からは，互いの石の配置などの静的局面情報や着手系列によって局面がどのように変化していったかなど，ゲームの進行に関する情報を全て再現することができる．また，囲碁は別名「手談」とも言われるように，その対局は着手を通じたコミュニケーションであるとみなすことができ，個々の着手にはこれ以外にもその局面における役割やプレイヤの意図などの様々な情報が内在されていると考えられる．個々の着手およびその系列が持つこのような情報を棋譜から抽出することは，そのゲームの内容を理解するためには欠かせない作業であり，また，結果として得られたこれらの情報は，局面認識や着手決定のためのパターン知識など幅広い利用が期待できる．

筆者らは，この問題に対して，自然言語処理とのアナロジーから，棋譜を個々の着手を符号化してできたテキスト（棋譜テキスト）であるとみなし，この「棋譜テキスト」に対して適切な言語モデルを作成し言語処理技術を適用するというアプローチを行なっている [2][3][4]．そして，棋譜テキストに対する言語モデルが，盤面認識や着手候補選出などゲームプレイシステムの中核として利用できるだけでなく，ゲーム記述言語と自然言語との相互変換を通じて，棋譜からの解説文生成や自然言語による棋譜データベース検索など，知的ゲームに関する自然言語システムへと幅広く応用することができ，ひいては，言語以外の分野における人間の言語的思考過程の解明へと繋がることを期待している．

自然言語テキストにおいて，単語はそれ自体で意味を担っており，単語同士が結合することによって句や文全体の意味が形成される．したがって，自然言語テキストとのアナロジーを考えて棋譜テキストを自然言語テキストと同様に取り扱うためには，棋譜中の個々の着手を単語とみなして，これに対してどのような特徴を捉えた符号化を行ない棋譜テキストを構成するかが重要な問題となる．筆者らがこれまで行なってきた棋譜からの定型手順知識獲得の研究においては，直前の相手方の着手との相対的な位置関係をもとにした着手符号化を行ない，文字 n-gram に基づく定型性の評価によって，これまでの固定窓を使う手法では獲得が難しかった長さの異なる様々な定型手順パターンを自動的に獲得できることを示した [2][3]．これは，相対的位置関係の系列が，着手列によって構成される石の一団の形状を適切に反映するような特徴を有していたためであると考えられる．しかし，棋譜解析によってゲーム内容を理解し，棋譜中に記述されている様々な情報を抽出することを考えた場合には，この他にも着手の特徴として記述しておくべき情報が存在する．

そこで本論文では，言語処理に基づく棋譜内容解析への出発点として，棋譜をその内容に応じて分類するというタスクを設定し，テキスト自動分類の手法を用いて棋譜中の着手の特徴抽出とそれを用いた棋譜の分類を行なう．

2 着手の符号化

囲碁では，チェスや将棋とちがって個々の石には先験的な役割が与えられておらず，周囲の状況や局面の進行状況に応じて着手された石の役割が定まる．また，プレイヤの着手意図や戦略などに関する情報は，ある時点の特定の着手だけではなく，一連の着手系列の中に存在していると考えることができる．したがって，個々の着手を単語になぞらえて 1 局の棋譜を棋譜テキスト化[1]する際には，このような着手および着手系列中に内在された情報が，テキスト中の語自体が持つ語彙情報として，あるいは，語と語の間の関係情報として適切に表現される必要がある．ここで，1 つの語が担う語彙情報を複雑にすればするほど，語と語の関係のルール（文法）を作成するのは困難になり，また，1

[1]棋譜テキストにおいては，1 手の着手に対して 1 つの符号が与えられる．これは，また，自然言語テキストにおける単語に相当するものであるとみなす．

つの語が担う語彙情報を単純にしすぎれば，その文法の記述能力が低下し十分な情報を表現することができなくなるといったトレードオフがある．

我々は，この問題に対して，まず，単純な語彙情報を持つ単語からなるテキストがどの程度の記述能力を持つのか（棋譜テキストに関していえば，着手に対して基本的で単純な特徴素を抽出して符号化された棋譜テキストがどのような情報を担っているのか）を明らかにする目的で，着手された石自体の役割を次の３つの側面から眺めて特徴づけることにする．

- 着点の盤面上の絶対位置
- 直前の相手方の着手と現在の着点との位置関係
- 着点の周囲の状況

これらの各々の観点に対して，石の形や筋といった囲碁の専門知識を必要としない基本的な特徴を用いて個々の着手を符号化する．

2.1 盤面上の絶対位置に関する符号化

囲碁では，盤上の位置を表わす表現として「星，天元，三々，小目，高目，隅，辺，中央，三線，四線，...」など様々な用語が存在する．これは即ち，着点がどこに位置するかという情報が，その着手の特徴として有効に利用できることを裏づけている．盤面上の位置情報の符号化にあたって，例えば着点 (17, 4) や (4, 3) などはいずれも「小目」であり，これらの着手には同じ符号を与えることが望ましい．したがって，着手に対して盤上での回転，鏡像に関する変換の影響を受けないような以下の符号化を行なう．

盤面上の絶対位置に関する符号化: E_{loc}

天元 (10, 10) を座標原点とした相対座標に基づいて符号化する．現在の着手の着点が (x, y) であるとき，着手符号 $c_{x',y'}$ は次式で与えられる．

$$\begin{cases} x' = \min(|x-10|, |y-10|) \\ y' = \max(|x-10|, |y-10|) \end{cases}$$

2.2 直前着手との位置関係による符号化

「相手方の着手に対してどのように応対するか」ということは，着手を行なう際に考慮すべき最も基本的な事柄である．激しい接近戦が繰りひろげられている状況では，当然，相手方の着手に接近した地点に着手される可能性が高く，また，部分戦が一段落した状況では，他方面の価値が高い箇所が着手候補となる．定石や手筋などといった定型的な着手系列知識を獲得する目的で筆者らが以前に行なった実験 [2][3] では，直前の相手方の着手との相対的な位置関係をもとにした以下のような着手符号化を行なった．以下で説明する符号化法は，着手列によって構成される石の一団の形状を，回転，移動，鏡像などの影響を除いて適切に反映するという特徴を有している．

直前着手との位置関係に関する符号化: E_{pre}

直前の相手方の着点を座標原点とした相対座標に基づいて符号化する．x 軸，y 軸の選択と正負の方向は，直前の着点が図 1 で示した各領域のどれに属するかに応じて決定する．この座標系の上で現在の着点の相対位置 (x, y) を求め，符号 $c_{x,y}$ を与える．

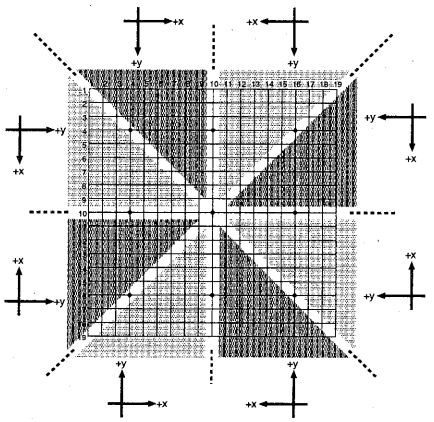

図 1: 領域と座標軸との対応

2.3 着点の周辺配石に関する符号化

石の形や筋といった概念は，着点とその周辺の石の配置状況によって定まる．清らは，着点を中心としたマンハッタン距離 4 以下の菱形領域内の石の配置を形パターンデータとして用いることにより，良い候補手が生成できると報告している [1]．そこで，我々もこれと同じ手法を用いて周辺配石に関する符号化を行なう．

着点の周辺配石に関する符号化: $E_{sur}(M)$

着点からマンハッタン距離 M 以内にある各格子点に対して，その点の状況（味方の石，相手の石，空点，盤外）に応じて 2 bit の符号を与え，これを範囲内の全点について連結したものをその着点の符号とする．

(印字可能文字からなるテキストとして処理するため，実際には，3 点分の 6bit を 1 バイトに伸張して符号化を行なっている．したがって，図 2 に示すマンハッタン距離 4 以内にある 40 個の格子点の状況は，14 バイトの符号によって表現される)

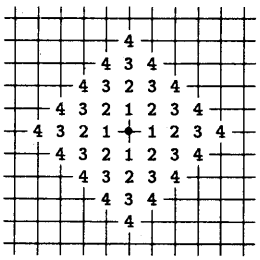

図 2: マンハッタン距離 4 以内の領域

3 棋譜テキストの分類法

自然言語テキストの検索や分類では，通常，テキストから特徴素として索引語 (index term) を抽出し，抽出された索引語集合を用いてテキストの内容を近似するという方法が取られる．ここで重要となるのは，テキストの内容を適切に反映する索引語集合を漏らさずに抽出することであるが，索引語としてどのような単位のものをどういった基準で抽出するかに関しては，対象となるテキストの特徴に応じていくつかの手法が提案されている [6][7][10]．

そして，得られた索引語集合に基づいて文書の検索，分類を行なう手法としては，ベクトル空間モデル [8][11] や確率モデル [9][12] などが代表的な手法として広く用いられている．

以下では，まず，本研究の棋譜テキスト分類手法として用いたベクトル空間法について説明する．

3.1 ベクトル空間法

テキスト集合を $D = \{d_1, d_2, \ldots, d_N\}$，索引語集合を $T = \{t_1, t_2, \ldots, t_n\}$ としたとき，テキスト $d\ (\in D)$ の特徴は，このテキストにおける各索引語の重要度を要素とする次のベクトル V_d で表現される．

$$V_d = (w_1^d, w_2^d, \ldots, w_n^d)$$

ここで w_i^d は，索引語 t_i のテキスト d に対する重要度を表わす．

そして，テキスト d と d' との類似性を，特徴ベクトル V_d と $V_{d'}$ の類似度 $sim(V_d, V_{d'})$ を用いて評価することによってテキストの検索や分類が行なわれる．テキスト分類で用いられるベクトル間の類似度としては，以下の余弦測度が一般的である．

$$sim(V_d, V_{d'}) = \frac{\sum_{k=1}^{n}(w_k^d \cdot w_k^{d'})}{\sqrt{\sum_{k=1}^{n}(w_k^d)^2 \cdot \sum_{k=1}^{n}(w_k^{d'})^2}}$$

3.2 特徴素の選出

西野は，[6] において日本語テキスト分類における索引語としてどのような単位が有効であるかを，文字，単語，句など様々な単位に対する系列に対して実験的に検証し，語の出現頻度情報などを用いて選別することにより，内容語

となる漢字 bigram や単語 bigram が有効な索引語として使用できると報告している.

棋譜テキストでは，着手符号化により個々の着手に対して一定の意味を担った内容語に相当する符号が与えられたとみなすことができるので，特徴素となる索引語候補としては，同様に n-gram (n=1, 2, 3) 文字列を採用することにするが，棋譜テキスト中にはこのような索引語候補は多数存在するので，この中から，棋譜の分類に寄与する可能性の高い索引語を $TF \cdot IDF$ 法を用いて以下の 4 通りの方法で選出する.

1. テキスト集合 D に対して分類カテゴリが $c_1, c_2, \ldots c_m$ として与えられているとする. $(c_i \subseteq D,\ D = c_1 \cup c_2 \cup \ldots \cup c_m)$
   そして，索引語候補 t のテキスト d 中での出現頻度 (文書内頻度) を $TF(t,d)$ ，t が出現するテキストを含むカテゴリ数を $CF(t)$，t が出現するテキストを含む文書数を $DF(t)$，として，以下の式で与えられる値をその索引語候補の重要度とする.

$$(\sum_{d \in D} TF(t,d)) * (\log(\frac{|D|}{CF(t)}) + 1)$$

2. この重要度は，多くのカテゴリに出現するカテゴリ特定性の小さい語や出現頻度の低い語に対して小さな値を取る．そこで，この値の大きい順に以下の基準にしたがって索引語を抽出する.

   (選出法 1) 無条件に抽出.

   (選出法 2) $CF(t) \leq \alpha$ (αは定数) となる t を抽出.

   (選出法 3) $\frac{1}{100} \leq \frac{DF(t)}{|D|} \leq \frac{1}{10}$ となる t を抽出.

   (選出法 4) 他の $TF(t,c)$ と比べて値が突出している $TF(t,c')$ が存在する t を抽出.

3.3 カテゴリベクトルによる棋譜テキストの分類

前節の手法により選出された索引語集合 $T = \{t_1, t_2, \ldots, t_n\}$ を用いて，テキスト $d\ (\in D)$ の特徴ベクトル V_d を以下のように構成する.

$$V_d = (TF(t_1,d), TF(t_2,d), \ldots, TF(t_n,d))$$

そして，カテゴリ c の特徴ベクトルは，そのカテゴリに属するテキストの特徴ベクトルの重心とする.

$$V_c = \frac{1}{|c|} \sum_{d \in c} V_d$$

分類対象テキスト d が与えられたとき，d は $sim(V_d, V_c)$ を最大にするカテゴリ c に分類される.

4 棋譜の分類実験

ここで述べた棋譜テキストの分類手法がどの程度有効であるか，また，どのような着手符号化と特徴素の選出法の組合せがマッチするのかを調べるために，棋譜テキストの分類実験を行なった.

4.1 分類カテゴリの設定

最終的な目標は「棋譜をそのゲームの内容（例えば,「激しい戦いの碁」や「大模様の碁」,「平明な碁」... など）に応じて分類すること」であるが，大量の棋譜データに対してあらかじめ内容を表わすタグを設定するためには高度な専門的知識が必要とされ，また，内容自体の分類カテゴリとしてどのようはものを設定しておくかということに関しても明確な基準は存在しない．そこで今回は，ゲーム内容に深く関係し，また，自動的に設定できる「対局者」をカテゴリとして採用することにする.

棋士にはその人独自の棋風というものがあり，これは，当然ゲーム内容に反映される．例えば,「黒：武宮九段 対 白：趙棋聖」の対局では，第 9 期棋聖戦挑戦手合の対局にみられるように，典型的な大模様対実利という対決となる可能性が高い．したがって,「対局者」をカテゴリとする分類に有効な特徴素は，同様に棋譜の内容に関する分類にも有効に働くことが期待できる.

4.2 使用データ

日本棋院「棋譜データ集96」CD-ROMに収録されている33,685棋譜に対して，黒番の対局者と白番の対局者をそれぞれカテゴリとして割り当てることによって，全体として1,097のカテゴリ（黒551カテゴリ，白546カテゴリ）が得られた．しかし，各カテゴリに属する棋譜数はカテゴリによって大きく偏っており（最小で1棋譜，最大で749棋譜），このままでは分類のためのカテゴリサイズのバランスが悪いため，この中からカテゴリ中の棋譜数が100以上となるカテゴリのみを採用した．そして，各カテゴリに対して，学習用棋譜データとして50個，テスト用棋譜データとして10個の棋譜をランダムに選出した．こうして，最終的に得られた以下のカテゴリと棋譜を分類に使用した．

カテゴリ数	222	(黒114, 白108)
学習用棋譜	10,044	
テスト用棋譜	2,188	

4.3 実験結果

まず，学習用棋譜データを E_{loc}, E_{pre}, $E_{sur}(1)$, $E_{sur}(2)$, $E_{sur}(3)$, $E_{sur}(4)$ の6通りの符号化法を用いて符号化し，その各々に対して初手から100手までの範囲内でn-gram (n = 1, 2, 3) を作成した．棋譜テキストは，両対局者の交互着手からなる符号系列であるが，分類カテゴリを「対局者」としたことを考慮すると，1手おきに取った一方の対局者のみの着手系列も特徴素として利用できるのではないかと考えられるため，1手おきの系列に対するn-gram (n = 2, 3) も作成した．各符号化法に対するn-gramの統計情報を表1に示す．

次に，このn-gramを索引語候補として，3.2節に示した4通りの選出法によって，1000語を上限として索引語を抽出した．ここで，(方法2) においては $\alpha = 0.8$ とし，(方法4) では $TF(t,c) > 3\sigma$ となるカテゴリ数を用いて抽出の可否を判定した．

選出された各々の索引語集合を用いて棋譜を特徴ベクトルとして表現し，学習用データ (closed data)，テスト用データ (open data) の各々に対して分類実験を行なった．主な結果を図3〜図10に示す．図3, 5, 7, 9がテスト用棋譜の分類結果，図4, 6, 8, 10が学習用棋譜データの分類結果である．グラフの横軸は分類カテゴリの順位，縦軸は正解率 (%) を表わし，その順位以内に正解が出現する累積正解率のグラフとなっている．テストデータに対して最も良い正解率が得られたものは，図3の「$E_{sur}(4)$ 符号化，1手おきの同色系列に対する3-gram，索引語選出法2」の組合せで，10位までの累積正解率が約47 (%)，正解カテゴリの平均順位は19.3位であった．これは，$E_{sur}(4)$ 符号化の符号長が14バイトと最も長く，符号中に着手が持つ役割などの情報が十分に反映されていたためであると思われる．E_{loc} と E_{pre} の比較に関しては，学習データの分類結果では E_{loc} 符号化の方が若干 E_{pre} の結果を上まわった．このことは，対局者の棋風と盤上の着手位置との間には何らかの相関があることを示唆している．また，図8と図10の比較より，索引語の選出方法としては，「選出法2」が少数の索引語を選出する際に有効であることが確認できた．図9にある $E_{sur}(2)$ の結果からわかるように，マンハッタン距離2程度の大きさの領域の配石状況は，符号長は4バイトと E_{loc} や E_{pre} より長いにもかかわらず，対局者に関する分類においてはあまり有効でないことが示された．

5 おわりに

囲碁の棋譜を着手毎の着点が符号化されてできたテキストとみなし，n-gram文字列を特徴素とするベクトル空間法によって棋譜の分類（対局者の推定）を行なった．「着点の盤面上の絶対位置」，「直前の相手方の着手と現在の着点との位置関係」，「着点の周囲の状況」の3つの観点からの着手符号化法を比較した結果，単独の手法では「着点の周辺の状況」に基づく符号化法が最も良い分類精度が得られた．ここで得

符号化法	符号長	黒白の交互着手系列			同色の着手系列	
	byte	1-gram	2-gram	3-gram	2-gram	3-gram
E_{loc}	1	55	2993	89316	3004	101208
		55	2961	62934	2977	75442
E_{pre}	2	659	86434	439385	90142	499971
		632	49554	92785	50545	104488
$E_{sur}(1)$	2	205	11281	109940	12041	122896
		203	8116	46276	8680	52303
$E_{sur}(2)$	4	74211	421063	627383	495105	724188
		34324	65172	53968	67449	43554
$E_{sur}(3)$	8	412079	686116	785586	738715	833081
		66498	50729	38556	44680	29639
$E_{sur}(4)$	14	639033	776051	827505	803426	855768
		51283	36918	31275	32870	25434

上段：学習用棋譜テキスト中の異り数，下段：頻度2以上のものの異り数

表1: 各符号化法における n-gram 文字列数

られた特徴素を基にして，例えば,「ある特定の定石を含む棋譜が持つ特徴の解析」などゲームの内容に関する解析と分類を行なうことが今後の課題である.

参考文献

[1] 清愼一, 川嶋俊明：“記憶に基づく推論を使った囲碁プログラム「勝也」の試作”, ゲームプログラミングワークショップ'97, pp.115–122, 1997.

[2] 中村貞吾：“*n*-gram 統計を用いた棋譜データベースからの定型手順の獲得”, ゲームプログラミングワークショップ'97, pp.96–105, 1997.

[3] 中村貞吾, 梶山貴司：“着手記号列の出現頻度に基づく囲碁棋譜からの定型手順獲得”, 情報処理学会ゲーム情報学研究会報告 GI 1-15, pp.107–114, 1999.

[4] 梶山貴司, 中村貞吾：“囲碁の着手記号列に対する確率文法モデルの作成”, ゲームプログラミングワークショップ'99, pp.161–168, 1999.

[5] 西野文人：“テキスト分類のためのカテゴリ割り付け戦略”, 情報処理学会自然言語処理研究会報告 NL 106-3, pp.13–18, 1995.

[6] 西野文人：“日本語テキスト分類における特徴素抽出”, 情報処理学会自然言語処理研究会報告 NL 112-14, pp.95–102, 1996.

[7] 小川知也, 落合亮, 西野文人：“文書クラスタ判別のための特徴表現付与”, 言語処理学会第5回年次大会, pp.209–212, 1999.

[8] 徳永健伸, 岩山真：“重み付き IDF を用いた文書の自動分類について”, 情報処理学会自然言語処理研究会報告 NL 100-5, pp.33–40, 1994.

[9] 岩山真, 徳永健伸：“自動文書分類のための新しい確率モデル”, 情報処理学会情報学基礎研究会報告 FI 33-9, pp.47–52, 1994.

[10] 竹内晴彦, 岩坪秀一, 西野博二：“多変量解析によるキーワードの自動抽出と文献の自動分類”, 情報処理学会情報学基礎研究会報告 NL 54-2, pp.1–8, 1986.

[11] Salton G., McGill, M.J. : “Introduction to Modern Information Retrieval”, McGraw-Hill, 1983.

[12] Robertson,S.E. and Sparck Jones, K. : “Relevance weighting of search terms.”, Journal of the American Society for Information Science, Vol.27, pp.129–146, 1976.

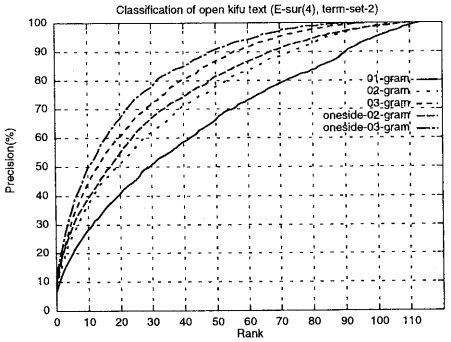

図 3: $E_{sur}(4)$, 選出 2, テスト

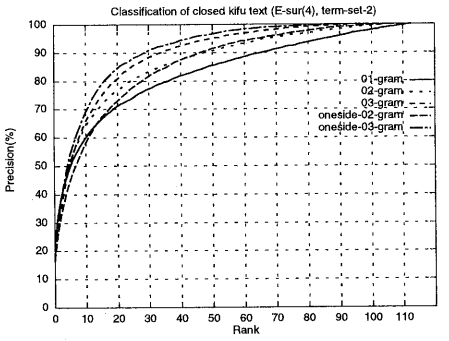

図 4: $E_{sur}(4)$, 選出 2, 学習

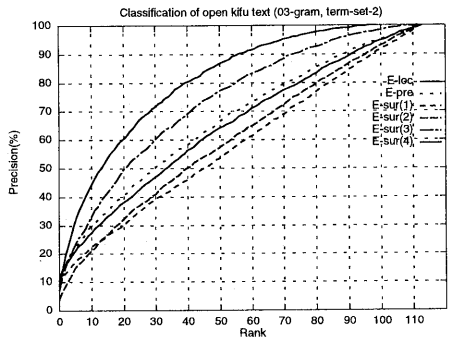

図 5: 3-gram, 選出 2, テスト

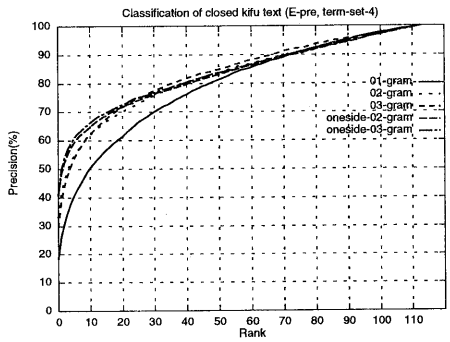

図 6: E_{pre}, 選出 4, 学習

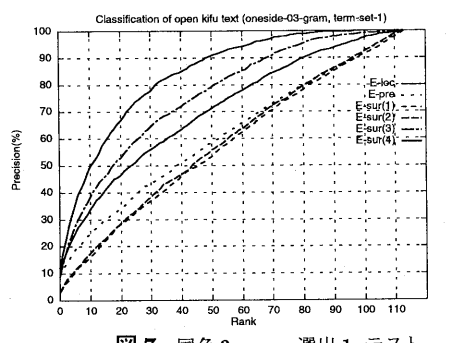

図 7: 同色 3-gram, 選出 1, テスト

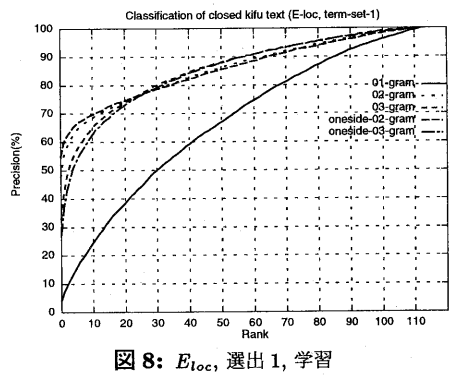

図 8: E_{loc}, 選出 1, 学習

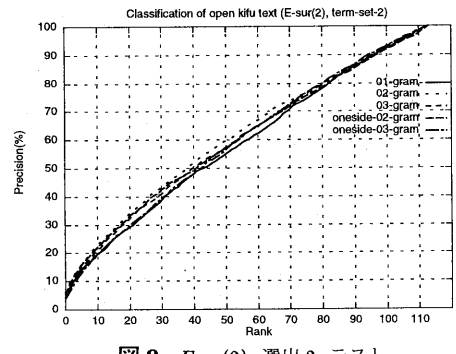

図 9: $E_{sur}(2)$, 選出 2, テスト

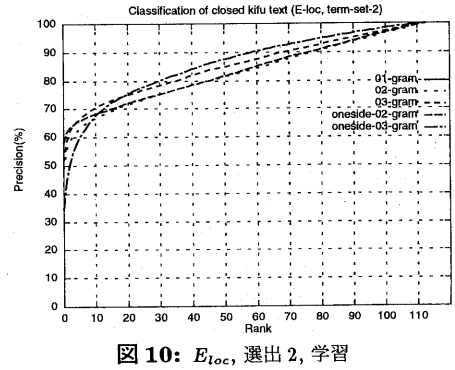

図 10: E_{loc}, 選出 2, 学習