人　工　知　能　106－16
（1996．10．31）

日本語文の類似性の学習

山本 公洋　内藤 昭三

NTT ソフトウェア研究所
〒180 東京都武蔵野市緑町 3-9-11

概要：
意味的な類似性に基づく正 / 負の日本語文事例集合に対して，語義および係り受け関係の 2 要因に関して，遺伝的アルゴリズムを用いた弁別学習システムを提案し，その性能評価を行なう．単語を有義語と連絡語に分類し，有義語のみが語義を持ち，正 / 負事例の使用状況においては，各単語の語義は一意であると仮定する．評価関数には，正事例に対する評価値の収束性，負事例に対する弁別性，係り受け解析木の大きさ，および係り受け解析木構成途中経過の類似性の 4 種を用い，多目的最適化を行なう．遺伝的アルゴリズムの適用では，異種族間交叉とビット出現頻度に基づく適応変異とを導入する．計算機シミュレーション結果を示す．

Learning Word Sense and Dependency from Synonymous Japanese Sentences

Kimihiro YAMAMOTO　Shozo NAITO
E-mail:{kimihiro, naito}@slab.ntt.jp

NTT Software Laboratories
9-11 Midori-Cho 3-Chome Musashino-Shi, Tokyo 180 Japan

Abstract:
This paper proposes a language acquision system based on the genetic algorithm. The system focuses on learning two linguistic aspects, i.e. word sense and dependecy from a set of positive/negative example sentences in terms of synonymity. The system learns the word senses and the dependencies from the surface information, i.e. words and their order for discriminating the synonymous sentences from non-synonymous ones.

We divide words into two categories: content and mediation words and suppose that only content words have their own sense and the sense is uniquely determined in the scope of given example sentences. We formalize the acquision problem as a multi-objective optimization, and applies the genetic algorithm to the problem. The system uses four criteria: semantic equality of the positive sentences, semantic discrimination between the positive and negative sentences, size of the dependency analysis tree of the example sentences and the similarity of the positive sentence analysing process. Genetic algorithm employs new methods: inter-tribe crossover and adaptive mutation depending on bit frequency.

1 はじめに

本稿では，意味的な類似性に基づく正／負の日本語文事例集合の表層的情報（単語・語順）から，語義・係り受け関係を学習するシステムについて述べる．同義文や言い換えなど，意味が等しい日本語文を類似文と呼ぶことにする．語義・係り受け関係に着目して日本語文の類似性を判定する．日本語文類似性判定を題材に，語義・係り受け関係の定義及び学習方法を考察する．

語義・係り受け関係は多義であり，文脈に依存する [1][4][5][7]．従来から，文脈情報を利用した語義・係り受け関係の多義選択方法が検討された．精度向上のためにモデルや規則の詳細化が要求された．その結果，モデル数や規則数が爆発する傾向にあった．これに対し本稿では，学習システムが取り扱う日本語文を事例範囲に限定すれば，語義・係り受け関係は一意であると仮定する．範囲限定により多義性に対する選択問題を回避する．事例集合別に可能世界を分けることで文脈を実現する．事例から語義・係り受け関係を学習する方法を提案し，語義・係り受け関係の文脈依存性に対処する．本稿では，一義性の観点から語義・係り受け関係を見直す．また，語義・係り受け関係の学習方法を提案する．

以下，第2章では，取り扱う新聞記事を示す．第3章では，語義・係り受け関係の一義性を検証する．第4章では，有義語と連絡語を定義する．第5章では，語義と係り受け関係の表現形式を提案する．第6章では，語義と係り受け関係の学習方法を考察し，多目的最適化問題と見立てる．第7章では，新しい遺伝的アルゴリズムを提案する．第8章では，計算機シミュレーション結果を示す．

類似文01 都市博中止を最終決断。
類似文02 都市博の中止が決まったことで、
類似文03 …都市博覧会の中止を決めたことについて「
類似文04 世界都市博覧会の中止決定について「
類似文05 …都市博覧会の中止を決定したことに関して「
類似文06 世界都市博覧会の中止決定に関し「
類似文07 …博覧会の中止を決断したことを受けて、
類似文08 …博覧会の中止を発表するにあたっての青島…
類似文09 都市博の中止を決断したからには、
類似文10 …博覧会の開催中止を最終決断したことについて、
類似文11 …博覧会の中止を決めた経緯を説明する。
類似文12 都市博中止決定について自民党内には「
類似文13 …博覧会の中止を最終決断したことに対して、
類似文14 …知事は世界都市博覧会の中止を決定したが、
類似文15 都市博の中止は決断したものの難しい撤退作業…

図 1: 類似文

2 取り扱う新聞記事

1995年4月1日～6月30日までの日本経済新聞（朝・夕刊，地方経済面含む）から，「都市博覧会（もしくは「都市博」）」と「中止（もしくは「中止する」）」を含む記事169件をキーワード検索する．次に，被験者1名が記事169件を読み，都市博覧会中止決定に関連深いと思われる記事48件を抽出する．記事48件を読点，句点，括弧で区切り，文に分割する．そして，「都市博覧会」と「中止」を含む101文を抽出する．さらに，被験者1名が101文を読み，類似すると思われる50文を抽出する．この50文を類似文として以下の考察を進める．類似文（一部）を図1に示す．類似文に含まれる10品詞，84種類，534語を図2に示す．数値は前者が単語出現回数，後者が単語の出現する類似文数を示す．

品詞	単語（語の数 / 文の数）
名詞	青島（幸夫）(24/24)，一夜(1/1)， 5日(1/1)，開催(8/8)，川崎市(1/1)， 幹部(1/1)，記者(1/1)，経緯(1/1)， 決断(2/2)，決定(12/12)，好意(1/1)， 31日(3/3)，午後(3/3)，午前(1/1)， こと(15/15)，コメント(1/1)，最終(7/7)， 自民党(1/1)，正式(4/4)， （世界）都市博覧会(37/37)，先月末(1/1)， 前日(1/1)，高橋清市長(1/1)，ため(1/1)， 中止(50/50)，知事(25/25)，1日(1/1)， 次(1/1)，撤去作業(1/1)，都(2/2)， 東京（都）(22/22)，東京都議会(1/1)， 東北地方(1/1)，通り(1/1)，都港湾局(1/1)， 都市博(13/13)，都市博特別委員会(1/1)， 内(1/1)，波紋(1/1)，2日(1/1)， 本番(1/1)，丸坊主(1/1)，論評(1/1)
動詞	明ける(1/1)，あたる(1/1)，いる(1/1)， 受ける(7/7)，会見する(1/1)，関する(2/2)， 決まる(2/2)，決める(15/15)，下す(1/1)， 決断する(12/12)，決定する(5/5)， 説明する(2/2)，対する(3/3)， 展開する(1/1)，なる(1/1)，発表する(1/1)， 開く(1/1)，表明する(1/1)，広げる(1/1)， 触れる(1/1)，迎える(1/1)
連体形容詞	難しい(1/1)
連体形容動詞	的な(1/1)
助動詞	た(39/36)，れる(1/1)
連体詞	ある(1/1)
格助詞	が(24/23)，から(1/1)，で(7/7)， に(15/14)，について(7/7)，によって(2/2)， は(9/9)，を(40/33)
連体助詞	の(54/42)
接続助詞	が(5/5)，から(1/1)，て(4/4)
連語	ての(1/1)，には(2/2)，にも(1/1)， ものの(1/1)

図 2: 取り扱う単語

3 一義性の検証

3.1 語義の一義性

類似文03に含まれる動詞「決める」は，類似文05に含まれる動詞「決定する」と置換可能である．置換可能な単語の語義は等しいと考えられる．単語の置換可能性を調べることにより，単語と語義との写像関係を検証した．

「決める」が出現する類似文は15文ある．この15文について調べた結果，全ての類似文において「決める」は「決定する」で置換可能であった．任意の単語について，その単語が複数回出現する場合，それら全ての単語が他の，ある1種類の単語で置換可能であれば，着目する単語の語義は一義に定まると考えられる．

類似文に含まれる名詞と動詞のうち複数回出現するものについて，上記の同一単語置換可能性を検証した．この結果，全ての名詞と動詞が同一単語置換可能であった．このことより，類似文範囲において，名詞と動詞の語義は一義に定まると考えられる．また出現回数が少なく厳密な検証を行えないが，文法的特徴の類似性より連体形容詞，連体形容動詞，助動詞，連体詞についても語義は一義に定まると考えられる．

一方，類似文14に含まれる格助詞「は」は格助詞「が」と置換可能だが，格助詞「を」とは置換できない．また，類似文15に含まれる「は」は「を」と置換可能だが，「が」とは置換できない．類似文範囲内に限定したにもかかわらず，格助詞の語義は一義に定まらない．連体助詞や接続助詞，連語についても，語義は一義に定まらないと考えられる．

3.2 係り受け関係の一義性

単語の修飾・被修飾関係を係り受け関係と呼ぶことにする．修飾語を係りの語，被修飾語を受けの語と呼ぶことにする．類似文に含まれる名詞「中止」に着目する．「中止」の係り先を第1受けの語と呼ぶことにする．また，第1受けの語の係り先を第2受けの語と呼ぶことにする．第1受けの語と第2受けの語を図3に示す．

係りの語	第1受けの語	第2受けの語	個数
中止	決断	—	1
	決定	—	9
	が	決まる	2
		決定する	1
	について	—	1
	によって	—	2
	を	決める	12
		決断する	12
		決定する	4
		発表する	1
		表明する	1
	の	決断	1
		決定	3
			合計 50

図 3: 係り受け関係

先の置換可能性検証より，名詞「決断」と「決定」の語義は等しい．また，動詞「決まる」と「決める」，「決断する」，「決定する」，「発表する」，「表明する」の語義は等しい．連体助詞もしくは格助詞の有無を無視すれば，「中止」は同じ語義を持つ単語へと係る．格助詞や連体助詞の有無を無視できるならば，任意の単語について，その係り先の語義は一義に定まると考えられる．

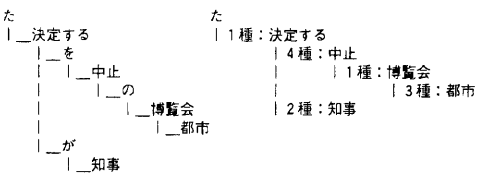
4 有義語と連絡語

格助詞，連体助詞，接続助詞，連語の語義は一義に定まらない．市販の辞書における格助詞等の意味数は，他品詞と比べ多い．また，格助詞等は単語種類が少なく，しかし文中における出現頻度が高い．そこで，格助詞等は語義を持たない，日本語文の構造を規定する記号と考える．格助詞，連体助詞，接続助詞，連語を以後，連絡語と呼ぶことにする．名詞，動詞，連体形容詞，連体形容動詞，助動詞，連体詞を有義語と呼ぶことにする．有義語の半順序関係として係り受け関係を再定義する（図4参照）．

5 表現形式

5.1 係り受け関係の表現

係り受け関係を係りの語，つなぎの語，受けの語，係種別の組で表現する（図4参照）．係りの語と受けの語は有義語とする．つなぎの語は3種類の値で表現する．つなぎの語を用いない場合をnで表現する．連絡語を用いる場合は連絡語を値とする．つなぎの語を用いても用いなくても良い場合は*で表現する．3種類の値を使い分け，表層の多様性を吸収する．



単語間の係り受け関係　　有義語間の係り受け関係

係りの語	つなぎの語	受けの語	係種別
知事	が	決定する	2種
都市	n	博覧会	3種
博覧会	*	中止	1種
中止	を	決定する	4種
決定する	n	た	1種

図 4: 係り受け関係

また，日本語文が有する一文一格の原則を考慮，係り受け関係の自由度を狭めて学習効率を上げるために格関係を導入する．格関係を含む係り受け関係を係種別と呼ぶことにする．全ての有義語に対し1～4種の係種別を設定する．1～4種の区別は各単語に依存し，変化するものとする．

5.2 語義の表現

単語ごとに用意する4種の係種別の役割は単語毎に変化する．4種の係種別総和を単語の語義とする．日本語文の意味（語義とその係り受け関係）を数値へと変換する．文字種類数が少なく文法構造が単純な数値表現を用いることにより，文の類似度に関する近似照合を可能にする．まず，4種の係種別ごとに演算子と数値の組からなる数式を対応させる．演算子の値は＋（加算）もしくは×（乗算）とする．数値は2～32の範囲の値とする．

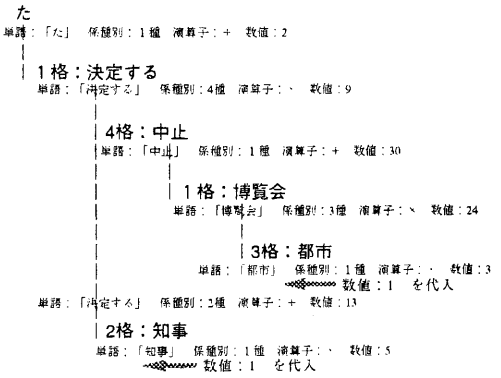


図 5: 語義の表現法

日本語文を有義語とその係り受け関係へ写像した結果を係り受け解析木と呼ぶことにする．係り受け解析木を数値演算過程と見なす．葉ノードに相当する語義に数値 1 を代入する．語義は演算子と数値を用いて計算し，返値を出力する．葉ノードから根ノードに向かって数値計算を繰り返す．係りの語である語義の返値を受けの語の引数とする．受けの語へ複数の係りの語が係る場合，係種別毎に数値計算を行ない，その総和を返値とする．根ノードに相当する語義の返値を文の意味値と呼ぶことにする．図5に計算例を示す．意味値を比較することで，文の類似性を判定する．意味値が等しいとき，2つの文は類似すると判定することにする．

6 語義と係り受け関係の学習

6.1 前提

類似する日本語文の表層的情報（単語・語順）から，語義・係り受け関係を学習する．類似文に共通する単語やその語順を手がかりに，語義・係り受け関係を学習する．

類似文 都市 博 の 中止 を 決める た

類似文 都市 博 の 中止 を 決定する た

上記2つの類似文では「都市」，「博」，「の」，「中止」，「を」，「た」の6語が共通する．上記類似文より，動詞「決める」と「決定する」の語義が等しいことが学習できる．しかし，単語の違いや語順の違いなど，類似文では複数の相違点が混在する．語義・係り受け関係は個別に学習できない．これに対し本稿では，意味値の差分（数値距離）を意味の「近さ」に見たてる．最適化手法を用い，語義・係り受け関係を学習する．

6.2 評価関数

類似文の正事例と負事例より，語義・係り受け関係を学習する．正事例の意味値が等しくなるように，語義・係り受け関係を学習する．また正事例と負事例の意味値が異なるように，語義・係り受け関係を学習する．前者を収束性，後者を弁別性と呼ぶことにする．意味値間の数値距離から収束性と弁別性を測定する（図6参照）．

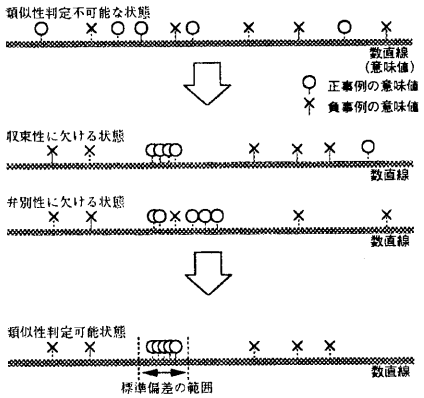


図 6: 収束性と弁別性

$$a = \frac{b}{\overline{x}},\ b^2 = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m}$$

正事例の意味値の標準偏差を計算し，収束性を測定する．a は収束度，b は標準偏差，x_i は正事例の意味値，$\overline{x}$ は正事例の意味値の平均値，m は正事例の個数を示す．標準偏差 b は平均値 $\overline{x}$ の影響を受け易い．従って，収束度で収束性を測定する．正事例の意味値が近いほど，収束度は小さくなる．

$$c = \frac{\sum_{j=1}^{n} d_j}{n},\ d_j = \begin{cases} 0 & \text{if } |y_j - \overline{x}| > \sigma \\ 1 & \text{otherwise} \end{cases}$$

正事例の意味値が分散する範囲に負事例の意味値が含まれてしまう度合を計算し，弁別性を測定する．標準偏差の範囲を，正事例の意味値が分散する範囲とする．c は弁別率，y_j は負事例の意味値，n は負事例の個数を示す．σ と $\overline{x}$ は評価関数1と同じ値を示す．正事例と負事例の意味値が離れれば，弁別率は小さくなる．

また，収束度や弁別率などの1次的尺度に対し，2種類の2次的尺度を導入する．係り受け関係を用いて正／負事例から係り受け解析木が生成される．大きな係り受け解析木が生成できれば，弁別性に優れていると考えられるので，係り受け解析木の大きさ（ノード数）を測定する．また，意味値計算の途中経過（数値）が似ていれば，収束性に優れていると考えられる．このために，数値の出現回数の2乗の総和を測定する．合計4種類の評価関数を用いる．

4種類の評価関数を一次元的尺度で測定するのは難しい．本稿では，遺伝的アルゴリズムを用いて多目的最適化を行う．語義・係り受け関係をビット列で表す．n個のビット列を集団（個体群）で保持する．評価値に準ずる確率で，個体群から2個のビット列を選択する（選択淘汰）．選択した2個のビット列に交叉，突然変異の順番で2種類の遺伝操作を加え，新しいビット列を生成する．上記の選択淘汰と交叉，突然変異をn/2回繰り返し，新しい個体群を作る．新しい個体群を繰り返し生成しつつ，最適化を行なう．

7 独立並列方式

7.1 ビット列

語義・係り受け関係を別々にビット列化する．2種類の固定長ビット列を用いる（図7参照）．

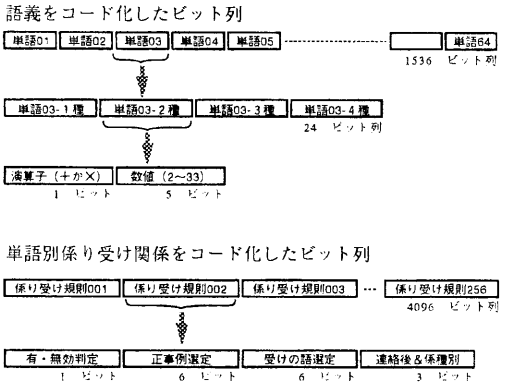


図 7: ビット列

ひとつの語義を6ビットで表す．先頭1ビットで演算子を表す．0の場合は＋（加算），1の場合は×（乗算）を表す．残り5ビットで数値を表す．システム辞書登録可能な有義語数を64語とする．それぞれの有義語は係種別：1～4種に相当する語義を持つ．語義を1536ビット列で表す．

ひとつの係り受け規則を16ビットで表す．先頭1ビットで係り受け関係の有・無効を表す．次の6ビットで復号の際に用いる正事例を表す．次の6ビットで受けの語を表す．残り3ビットで連絡語，係種別を表す．システム辞書登録可能な有義語64語それぞれについて，有義語を係りの語とした係り受け関係を4つ用意する．係り受け関係を4096ビット列で表す．

係り受け関係は以下の手順で復号する．まず，係り受け関係の有・無効を判定する．有効な場合のみ，残り15ビット列を復号する．システム入力の正事例集合から，係りの語を含む正事例を1個選定する．次に，選定した正事例から，受けの語を選定する．最後に，選定した正事例や受けの語を考慮しつつ，連絡語と係種別を設定する．

7.2 ＧＡによる同時並行探索

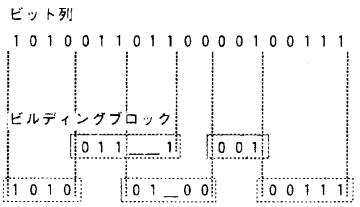


図 8: ビルディングブロック

学習対象であるビット列を部分に分割する．この部分をビルディングブロックと呼ぶ．以後，ブロックと略す．遺伝的アルゴリズムの利点は，この複数のブロックを同時並行探索できることであると考えている．任意のブロック学習過程が他のブロック学習過程に与える影響を取り除のぞく方法として以下，独立並列方式を提案する（図9参照）．

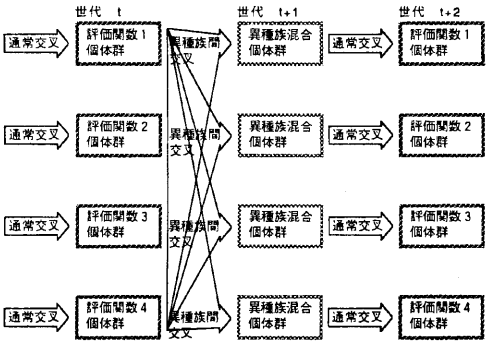


図 9: 独立並列方式

7.3 評価関数毎の選択淘汰

従来の多目的最適化ではトレード・オフの取り扱いが重要課題であった．複数の評価関数を一次元的尺度に統合していた[2][3]．これに対し独立並列方式では，ブロックの個別学習を重要課題と考える．解を構成するブロックの役割は多様であると考える．多様な役割を個別に評価するため，複数の評価関数が必要と考えている．

評価値の相互影響を取り除くため，複数の評価関数を個別に用いる．評価関数毎に個体群を準備する．各個体群は独立に学習を進める．但し，一定時間間隔で生じる移住世代において，個体群間でブロックを交換する．

St01 正事例と負事例を入力する.

St02 1組（2本）のビット列をn個ランダム発生させる.

St03 $i=0$, $j=0$ とする.

St04 $i=i+1$, $j=j+1$ とする. $i=$終了世代の場合, *St12* へ進む.

St05 n組の各ビット列について, *St05-01*～*04* を行う.

St05-01 ビット列から, 語義と係り受け関係を復号する.

St05-02 係り受け関係を用いる. 正・負事例から係り受け解析木を生成する. 事例数分の係り受け解析木が生成される.

St05-03 語義を用いる. 係り受け解析木から意味値を計算する.

St05-04 評価関数を用いる. 意味値を見比べ, 評価値を計算する.

St06 $j=$移住世代なら *St09* へ, さもなくば *St07* へ進む.

St07 *St07-01*～*03* をn/2回繰り返す.

St07-01 評価値に準ずる確率で, n組のビット列から2組のビット列を選択する（選択淘汰）.

St07-02 選択した2組のビット列を交叉させ, 新しいビット列を生成する（通常交叉）.

St07-03 新しいビット列に適応変異を加える（適応変異）.

St08 n組のビット列をn組の新しいビット列で置き換える. *St04* へ戻る.

St09 優良ビット列（複製）を他の個体群へ送り出す. また, 他の個体群の優良ビット列を受けとる.

St10 *St10-01*～*03* をn/2回繰り返す.

St10-01 優良ビット列からランダムに1組のビット列を選択するまた, 適応値に準ずる確率で, n組のビット列から1組のビット列を選択する（選択淘汰）.

St10-02 選択した2組のビット列を交叉させ, 新しいビット列を生成する（異種族間交叉）.

St10-03 新しいビット列に適応変異を加える（適応変異）.

St11 n組のビット列をn組の新しいビット列で置き換える. $j=0$ とする. *St04* へ戻る.

St12 終了する.

図 10: 学習アルゴリズム

7.4 異種族間交叉

標準的な遺伝的アルゴリズムは, 同一評価関数に基づき選ばれた2つのビット列を交叉させる. これを通常交叉と呼ぶことにする. これに対し独立並列方式では, 異なる評価関数に基づき選ばれた2つのビット列を交叉させる. これを異種族間交叉と呼ぶことにする. 独立並列方式では, 通常交叉と異種族間交叉との両方を用いる. 移住世代において異種族間交叉を用い, ブロックを交換する. 通常交叉, 異種族間交叉ともに2点交叉を用いることにする.

通常交叉だけでは, ある特定の目的のみに偏った解しか得られない [8]. 従来は複数の評価関数を一次元的尺度に統合することで, パレート最適解を学習した. 独立並列方式では, 適切なブロックを組み合わせることで, おのずと多目的最適解が得られると考えている. 従来は複数の評価関数を一次元的尺度に統合することで, パレート最適解の繁殖確率を強制的に高くしていた. 独立並列方式では, 異種族間交叉を用いることで, パレート最適解の保存確率を自然と高くしている. 図10に独立並列方式による語義・係り受け関係の学習アルゴリズムを示す.

7.5 出現頻度に基づく適応変異

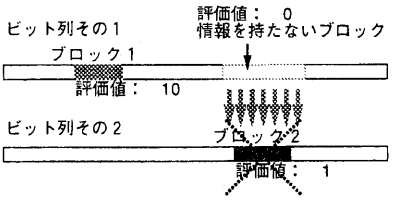


図 11: ビルディングブロックの相互影響

突然変異確率を動的に変化させる. 局所解に収束したとき, 突然変異確率を高くして個体群内の多様性を保つ.

Whitleyが提案した適応変異 [9] では, 交叉により生成した2つのビット列のハミング距離を測定し, ハミング距離が近いほど高い突然変異確率を与えた. しかし, 任意のブロック学習過程が他のブロック学習過程に与える影響を取り除けない図11では, ブロック2評価値よりもロック1評価値が高い. 個体群内部にブロック1を含むビット列が繁殖する. ブロック1のビット長が短い場合, 先の適応変異で多様性を保つのは難しい. ブロック1を含むビット列がブロック2の学習を阻害する.

これに対し並列独立方式では, 遺伝子座単位でビット出現頻度を測定する. 同じビットが頻繁に出現する場合, その遺伝子座の突然変異確率を高くする. ビット列1個について, 遺伝子座毎のビット出現頻度をカウントするために, 同じ長さの自然数列を用意する. 自然数列の初期値は0とする. 交叉により生成した2つのビット列を比較, 同じ遺伝子座にあるビットが等しい場合, カウンタを1増やす. 同じ遺伝子座にあるビットが異なる場合, カウンタを1減らす. 但し, カウンタの下限は0とする. カウンタの大きさに準ずる確率でビットを反転させる. ビットを反転させた場合, カウンタは初期値に戻す.

8 実験

8.1 パラメータ

正事例9個と負事例11個をシステムに与える（図12参照）. 4種類の遺伝的アルゴリズムの探索能力を比較する. 手法1: Fonsecaの手法 [2], 手法2: Schafferの提案したVEGA (Vector Evaluated Genetic Algorithm) 法 [8], 手法3: 独立並列方式その1, 手法4: 独立並列方式その2を比較する.

手法1は, 4つの評価関数を一次元的尺度に統合する. その他では, 4つの評価関数を個別に用いる. 手法1と手法2では, 通常交叉のみを用いる. 手法3と手法4では, 通常交叉＋異種族間交叉を用いる. 手法3では, Whitleyの提案した適応変異を用いる. その他では, 出現頻度に基づく適応変異を用いる. 手法1ではパレート最適解を重視する. 世代交代の際, 新・旧あわせて2n組のビット列から, パレート

正事例01 世界 都市 博覧会 の 中止 を 決定する た
正事例02 世界 都市 博 の 中止 を 決定する た
正事例03 世界 都市 博 中止 を 決定する た
正事例04 都市 博 の 中止 を 決定する た
正事例05 都市 博 中止 を 決定する た
正事例06 都市 博覧会 中止 を 決断する た
正事例07 都市 博 中止 を 決断する た
正事例08 都市 博覧会 の 中止 を 表明する た
正事例09 都市 博 を 中止する た

負事例01 中止 を 決定する た
負事例02 中止 を 決断する た
負事例03 中止する た
負事例04 世界 都市 博覧会 を 中止する か
負事例05 世界 都市 博覧会 を 中止する ない
負事例06 中止 を 公約する て いる 博覧会
負事例07 都市 博 の 中止 を 公約する て いる
負事例08 都市 博 中止 を 強調する た
負事例09 都市 博 中止 を 撤回する
負事例10 都市 博 中止 は 難しい
負事例11 都市 博 中止 を 決断する ない

図 12: 入力事例

最適なn組のビット列を選択し新しい個体群とする．いわゆるエリート戦略を用いる．

手法1，3，4で評価関数別に生成されるビット列は250組とする．手法2で統合評価関数により生成されるビット列は1000組とする．いずれの手法でも1世代において生成するビット列は1000組とする．世代交代数は200とする．全ての手法において $1000 \times 200 = 200000$ 回の試行を行なう．手法3と手法4では，3世代毎に移住世代を発生させる．

交叉確率は1.0とする．適応変異では突然変異確率の範囲を，0.005～0.05とする．手法1～4について，同じ実験を5回繰り返す．

8.2 探索能力

収束度が0.05以下で，かつ弁別率が0%の解を正解とする．各手法について同じ実験を5回繰り返す．正解の得られた回数を図13に示す．また，各実験で得られた収束度，弁別率の最小値（平均）を示す．図14に学習した語義・係り受け関係の例を示す．係り受け解析，意味解析で用いる語義・係り受け関係をアンダーラインで示す．

正事例について，人手で語義・係り受け関係を作成する．語義の等しい単語を等価関係で結ぶ．係りの語と受けの語を係り受け関係で結ぶ．人手で作成した語義・係り受け関係と学習結果を比較する．人手で作成した語義等価関係（もしくは係り受け関係）の個数を e とする．また，システムが学習できた語義等価関係（もしくは係り受け関係）の個数を f とする．達成率を e/f とする．語義，係り受け関係の達成率の最大値（平均）を図13にあわせて示す．

手法1では5回全部，収束度0.0弁別率0.18の局所解に収束した．正解数は0だった．語義と係り受け関係の達成率は低い．複数の評価関数を一次元的尺度に統合すると，ブロック等を適切に評価できなくなる．つまり，探索で利用可能な情報が失われると考えられる．また，極端なエリート戦略（パレート最適優先）のため，探索の多様性が失われると考えられる．手法2では正解数は0だった．語義と係り受け関係の達成率は低い．手法2では特定の目的のみに偏った解しか得られなかった．これに対し手法4では収束度と弁別率の両方に優れた語義・係り受け関係が得られた．異種族間交叉の有効性が確認できた．手法3では，正解数が少ない．しかし，語義と係り受け関係の達成率が高い．これは，正解の近傍が得られたことを示している．Whitleyの適応変異ではハミング距離が近いほど突然変異率が高くなる．近傍探索能力に欠ける．これに対し手法4では正解が得られた．これにより，出現頻度に基づく適応変異の有効性が確認できた．

	Fonseca	VEGA	独立並列方式	
			その1	その2
正解数	0/5	0/5	0/5	3/5
収束度 最小値平均	0.00	0.10	0.00	0.00
弁別率 最小値平均	0.18	0.09	0.00	0.00
語義 達成率 最大値平均	39.2%	33.7%	97.9%	97.2%
係り 達成率 最大値平均	25.7%	54.2%	100.0%	100.0%

図 13: 探索能力

8.3 類似性判定能力

正事例数を36個とする．負事例数を11個（実験1と同じ）とする．36個からランダムに10個の正事例を選択する．これを学習事例とする．残りをテスト事例とする．学習事例10個と負事例11個をシステムに与える．最優良ビット列から語義・係り受け関係を復号する．復号した語義・係り受け関係を用いて，テスト事例を正事例として判定できるか，検証する．同じ実験を5回繰り返す．テスト事例に対する正解率を図15に示す．学習事例とテスト事例とに対する語義・係り受け関係の達成率も示す．

先と同様弁別率が1.0で，かつ収束度最小のビット列を最優良ビット列とする．1回目と3回目の実験では，収束度が大きい．また，学習事例に対する語義・係り受け関係の達成度も低い．このため正解率が低いと考えられる．2回目の実験では，収束度が小さい．学習事例に対する達成度も高い．しかし，正解率は低い．標準偏差を閾値とする場合，正事例の意味値にばらつきがあると，僅差であってもいくつかの正事例は見落とされる．僅差のテスト事例を考慮した場合，正解率15/26となる．残りテスト事例11/26の大半は，学習事例に含まれない未知語を含んでいる．4回目の実験では，弁別率1.0のビット列を学習できなかった．未知語への対応能力がないことを考慮すれば，2回目と5回目において高い正解率が得られたと考えられる．

係りの語	つなぎの語	受けの語	係種別
世界	—	—	—
世界	—	—	—
世界	n	た	3種
世界	—	—	—
都市	n	博覧会	3種
都市	n	博	4種
都市	n	た	4種
都市	n	た	1種
博覧会	—	—	—
博覧会	n	表明する	1種
博覧会	—	—	—
博覧会	*	中止	3種
博	—	—	—
博	*	中止	3種
博	を	中止する	1種
博	—	—	—
中止	を	決定する	3種
中止	を	決断する	3種
中止	を	表明する	1種
中止	—	—	—
中止する	—	—	—
中止する	—	—	—
中止する	—	—	—
中止する	n	た	3種
決断する	n	た	3種
決断する	n	た	4種
決断する	n	た	4種
決断する	—	—	—
決定する	n	た	3種
決定する	n	た	4種
決定する	n	た	1種
決定する	—	—	—
た	—	—	—
た	—	—	—
た	—	—	—
た	—	—	—

有義語	係種別	演算子	数値
世界	1種	×	15
世界	2種	×	22
世界	3種	+	24
世界	4種	×	25
都市	1種	+	22
都市	2種	+	24
都市	3種	×	23
都市	4種	+	14
博覧会	1種	×	31
博覧会	2種	+	26
博覧会	3種	+	31
博覧会	4種	+	17
博	1種	+	32
博	2種	+	20
博	3種	×	30
博	4種	+	31
中止	1種	×	23
中止	2種	×	22
中止	3種	+	10
中止	4種	×	25
中止する	1種	+	22
中止する	2種	+	8
中止する	3種	+	6
中止する	4種	×	25
決断する	1種	+	10
決断する	2種	+	28
決断する	3種	+	14
決断する	4種	×	12
決定する	1種	+	12
決定する	2種	+	28
決定する	3種	+	14
決定する	4種	+	2
た	1種	×	10
た	2種	+	15
た	3種	×	26
た	4種	×	19

図 14: 学習結果

9 おわりに

本稿では，日本語文の類似性判定を題材として，語義・係り受け関係のあり様や表現法について考察した．事例研究に基づき，単語を有義語と連絡語に分類した．有義語に着目し，係り受け関係を再定義した．また，語義・係り受け関係の学習方法を提案した．多目的最適化問題と見なし，遺伝的アルゴリズムを適用した．異種族間交叉と出現頻度に基づく適応変異を用いる独立並列方式遺伝的アルゴリズムを提案した．計算機シミュレーションを通して，提案手法の有効性を検証した．

今回は一部の新聞記事についてのみ考察したが，一般性を損なわないシステムを構築できたと考えている．文脈依存性を考慮して語義や係り受け関係を構築できる．情報検索等に用いた場合，高い照合精度が期待できると考えている．独立並列方式が全ての最適化問題に有効とは考えていない．関数最適化問題などにも適用し，有効範囲を明らかにしたい．

	正解率	達成率			
		学習事例		テスト事例	
		語義	係り	語義	係り
1回目	8/26	37.8%	71.8%	20.6%	59.6%
2回目	10/26	95.7%	100.0%	62.3%	90.6%
3回目	13/26	75.9%	68.6%	46.5%	55.2%
4回目	-	-	-	-	-
5回目	15/26	97.0%	100.0%	75.6%	78.9%

図 15: 類似性判定能力

謝辞

本研究では，株式会社 日本経済新聞社の許諾を頂き，日経全文記事データベース 日本経済新聞 CD-ROM95 版を使用しました．感謝します．また，日頃よりご支援頂く NTT ソフトウェア研究所 広域コンピューティング研究部 市川晴久部長，伊藤正樹グループリーダー，ならびに知的ソフトウェア研究グループのみなさまに感謝します．

参考文献

[1] 青沢, 笹野, 高木: 曖昧性解消ルールを用いた依存構造解析, 人工知能学会全国大会 第 9 回論文集, pp.449-502(1995).

[2] Fonseca,C.M. and Fleming,P.J.: Genetic Algorithm for Multiobjective Optimization: Formulation, Discussion and Generalization, *Proc. of the Fifth Int. Conf. on Genetic Algorithms*, pp.416-423(1993).

[3] Goldberg,D.E.: *Genetic Algorithms in Search, Optimization, and Machine learning*, Addison-Wesley(1989).

[4] 黒橋, 長尾: 格フレーム選択における意味マーカと例文の有効性について, 情報処理学会 自然言語処理研究会, 91-NL-11, pp.79-86(1992).

[5] 那須川: 文脈を利用した曖昧性解消法, 人工知能学会全国大会 第 7 回論文集, pp.425-428(1993).

[6] 野村, 井佐原, 徳永, 中村: 情報ハイウェイ時代のテキスト情報への知的アクセス, 情報処理学会誌, Vol.37, No.1, pp.1-9(1996).

[7] 奥村: 自然言語の意味的曖昧性の解消法, 人工知能学会誌, Vol.10, No.3, pp.332-339(1995).

[8] Schaffer,J.D.: Multiple Objective Optimization with Vector Evaluated Gegetic Algorithm, *Proc. of the First Int. Conf. on Genetic Algorithms and Their Applications*, pp.93-100(1985).

[9] Whitley,D. and Hanson,T.: Optimizing Neural Networks Using Faster, More Accurate Genetic Search, *Proc. of the Third Int. Conf. on Genetic Alrithms*, pp.391-396(1989).