自 然 言 語 処 理 105－1
（1995．1．20）

コスト最小法形態素解析のコストルール作成実験

小松 英二　安原 宏
沖電気工業株式会社 マルチメディア研究所

コスト最小法形態素解析では、プログラムが選んだコスト最小解が正解となるように、事前に個々のパタンに与えられる基本的なコストを設定しておく必要がある。本稿の形態素解析は、解の選好用のヒューリスティックを用いており、コストはヒューリスティックの確からしさになっている。コストは統計的手法ではなく、数理計画法的手法により設定する。本稿では、簡単な名詞句を解析対象として、選好ルールが与えるコストの値を決定する方法の実験結果について報告する。

An Experiment to Construct Cost Evaluation Rules for Morphological Analyzer of Minimum Cost Method

Eiji Komatsu　Hiroshi Yasuhara
Oki Electric Industry Co. Ltd.

In Morphological Analyzer of Minimum-Cost Method, values of costs given to patterns should be decided before execursions to ensure that a solution with minimum cost will be the solution. In the method of this paper, costs are decided by numerical planning method. To evaluate this method, expriments are done for simple noun phrases.

1．はじめに

コスト最小法形態素解析は、全解を圧縮したグラフ表現（グラフスタック、圧縮統語森等と呼ばれる）の始点から終点までの経路のうち、コストの総和が最小になる経路を選択し、その経路にある単語列をコスト最小解として出力する解析方法である。コストの付与方法としては、２つの単語の組み合せに対して与える方法［久光 1990］［久光 1991］、３つの単語の組み合せに対して与える方法［Nagata 1994］等がある。また、コストの意味付けとしては、コストを推移確率、出現確率として捉える確率モデルと、コストをヒューリスティックの確からしさとして捉えるモデル［吉村 1989］等がある。

コスト最小法が正しく動作するためには、事前に、コスト最小解が正解になるように各パタンに付与するコストの値を設定しておくことが必要である。確率モデルに基づくコストの場合は大量のデータによる統計的処理によりコストを設定することができるが、ヒューリスティックに基づくモデルの場合、ヒューリスティックの組み合せに確からしさを設定することが必要である。本稿の形態素解析は２つの単語の組み合せに対して、ヒューリスティックによりコストを与えるモデルを用いている。コストを与える方法としては、コスト決定用のルール（以下、コストルールと呼ぶ）の確からしさを変数としておき、数理計画法の手法により変数の値を決定する方法を提案した［小松 1990］。本稿では、本方式に基づいて試作したコスト決定プログラムによる実験結果、及び、本方法の有効性について述べる。

2．コストルール

本システムの形態素解析のコストルールは、解の候補から正解を選ぶための選好ルールである。ルール同士は必ずしも排他的ではなく、複数のルールのコストの和が、ノード又はアークのコストとして用いられる。図２－１にコストルールの例を示す。コストルールのコストは、解析結果のグラフにコストを与えるものであるが、実質的には、最も適当なルールの組み合せを選ぶという処理になっている。ルールの与えるコストは、変数としておく。コストルールは、ノードにコストを付けるルールとアークにコストを付けるルールがある。ノードのコストは、アークのコストの特別な場合であるが、ルールを作り易くするために両方とも用いている。パタンは自由に記述できる。コストルールはヒューリスティッ

クである以上かなり恣意的なものであるが、仮に間違ったルールを作成した場合には、コストの値（確からしさ）を更新していくに従って、値の絶対値が小さくなっていき、事実上削除されることを前提としている。

3．コストルールの作成方法

本稿のコストルール作成方法は、誤りのコストの総和より正解のコストの総和が小さくなることから、コストの変数の制約式を作成・蓄積し、これらの制約式を満たす値を求めることにより、コストの値を制限していく手法である。コストルールの作成は、人間が判断して作成する必要がある。一方、作成したルールのコストの値については、かなりの程度まで自動的に決定することができる。コストの値の決定は、「実例」を用いて行う。図３－１に実例の例を示す。実例は、「原文」、「解析結果の誤り」、及び、「修正した正解」からなる。以下に、コストルール作成のフローを示す。

［１］トレーニング・データを形態素解析する。
［２］トレーニング・データの解析結果を正解と比較し、解析結果が間違っている文について、実例を作成する。
［３］実例の誤りと正解に付与されるコストを変数の一次式で表わし、正解の一次式が小さくなるような制約式（不等式）を作成する。（図３－２参照）
［４］作成した制約式を既存の制約式に追加する。
［５］すべての制約式（連立一次不等式）を満たす値を１組求める。
［６］制約式が解けたときは、コストルールのコストの値を更新して［８］へ移る。
［７］制約式が解けないときは、人間の判断により、以下の１つの処理を行い、［５］へ移る。
・新しいコストルールを作成し、すべての実例についての制約式を作成し直す。
・現在処理中の実例の追加をやめる。
・追加した実例に反する既存の実例を削除する。（制約式において判断する）
［８］処理を終了する。

普通名詞_接尾語 → X0 ；品詞の組み合せにより単語ヒューリスティック
普通名詞 → X1 ；品詞に関するヒューリスティック
２文字単語 → X2 ；文字数に関するヒューリスティック
漢字の単語_漢字の単語 → X3 ；字種に関するヒューリスティック

図2－1　コストルールの例

【原文】１０月６日。
【誤り】1(数詞)/０(数詞)/月(単位)//６(数詞)/日(普通名詞)/。(記号)
【正解】１０月(普通名詞)/６(数詞)/日(後置助数詞)/。(記号)

図3－1　実例の例

【誤り】：蚊(普通名詞)/帳(接尾語)/。(記号)　【正解】：蚊帳(普通名詞)/。(記号)
↓ コストルールを適用する。
X0 + 3X5 + X10 + X30 + X38 + X243 + X855 + X1535 + X1551 + X1563 + X1615 + X1631 + X1643 + X1651 + X1656
> X5 + X6 + X10 + X38 + X883 + X1551 + X1563 + X1631 + X1643
↓ 不等式を整理する。
X0 + 2X5 - X6 + X30 + X243 + X855 - X883 + X1651 + X1656 > 0

図3－2　制約式の作成過程

4．コストルールの作成実験

3節で述べた方法に基づいて、コスト決定プログラムを試作した。図4－1に実験システムの構成を示す。実験は、名詞句に関して行った。ここで名詞句とは、名詞、助詞、接頭語、接尾語、連体詞の連続とする。トレーニングデータ及び評価用データは、名詞句１００個程度のグループとした（付録１～４）。コストルールは、図2－1に示したような品詞、品詞の組み合せ、字種、文字数に関するルールの４種類のヒューリスティック、約１７００ルールを初期状態として用意した。以下に、実験の手順を示す。

［1］トレーニングデータのグループを１つ選ぶ。
［2］3節のコストルール作成処理を行う。
［3］トレーニング・データの正解率が１００％ならば［7］へ、１００％でなければ［4］へ移る。
［4］評価用データを形態素解析し、正解率を求める。
［5］実験を続ける場合は、次のトレーニングデータを選び、［2］へ移る。

図4－2にルール作成過程におけるトレーニング・データ及び評価用データの解析成功率を示す。図では、3つのトレーニング・データについての結果を示した。第1グループでは、7個のルールを追加した。また、第2グループ、第3グループではルールの追加はない。評価データの解析成功率は、第1グループの終了時点で８７．６％、第2グループの終了時点で９３．６％、第3グループの終了時点で９４．６％だった。制約式が解けないため追加せずに捨てた実例は、第1グループで0個、第2グループで3個、第3グループで3個あった。これらは、サ変名詞と普通名詞、固有名詞と普通名詞、助詞の区別に関するものであり、見出しをパタンに入れないかぎり解決しないものである。表4－3に［4］で作成した実例の数を示す。なお、効率化のため、［4］について、最初の数段階では、一部の誤りに対する実例しか追加していない。

本実験の範囲では、図4－2に示すように、ルール作成は順調に進み、また、解析率が１００％に近くなるまでのコストルール作成処理の回数も現実的な範囲に収まっており、作成の効率、解析精度共に有効であると考えられる。本稿の実験で用いた文は比較的簡単であり、より複雑かつ大規模なトレーニングデータでの実験が必要であるが、形態素解析が用いている正規文法では隣接する単語同士の接続しか考えないことから、文が複雑になった場合でも、本実験の結果かなりの程度反映されると予想している。なお、不等式が矛盾して捨てざるを得ない実例があるため、解析精度が１００％になるまでルール作成処理を続けてはいない。

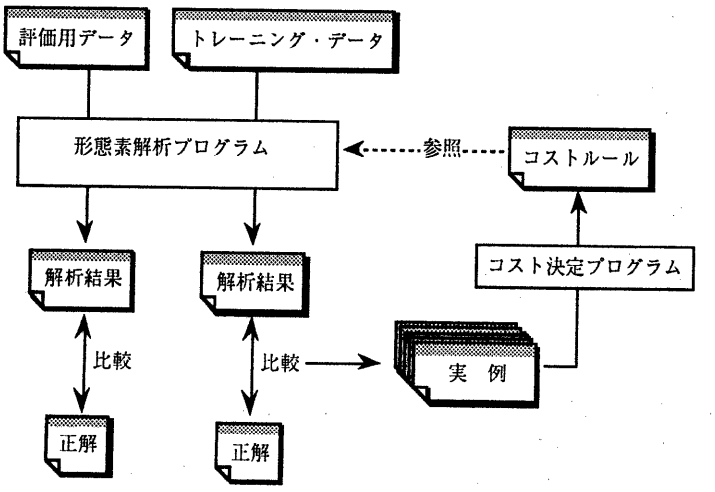

図4－1　実験システムの構成

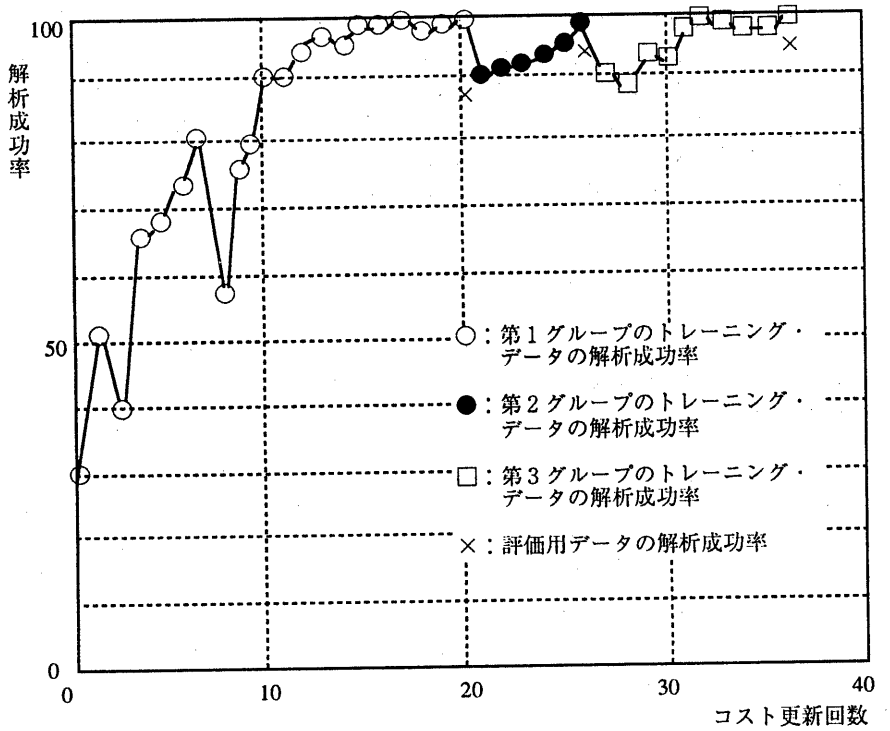

図4－2　トレーニング・データ及び評価用データの解析成功率

表4－3　作成した実例数

第1グループ

更新回数	実例数		更新回数	実例数	
	追加	累積		追加	累積
0	4	4	10	29	242
1	5	9	11	13	255
2	3	12	12	5	260
3	3	15	13	2	262
4	1	16	14	5	267
5	4	20	15	2	269
6	7	27	16	3	272
7	74	101	17	1	273
8	52	153	18	1	274
9	60	213	19	1	275

第2グループ

更新回数	実例数	
	追加	累積
0	20	295
1	4	299
2	5	304
3	4	308
4	2	310
5	3	313

第3グループ

更新回数	実例数	
	追加	累積
0	18	18
1	23	41
2	17	58
3	8	66
4	3	69
5	3	72
6	3	75
7	4	79
8	6	85
9	4	89

5．おわりに

本稿では、簡単な名詞句を用いて、コスト設定方式の有効性を示した。ルール自体の作り方、複雑な文、大規模なデータによる実験、不適当なルールを作ってしまった場合の副作用等が残されている課題である。本方式では、人手の介在が多いが、最初に充分なコストルールを与えておく、或は、制約式が解けない場合は無条件に実例を捨てて別途追加処理をする等の工夫により、自動化を進めることが可能

である。本稿の前提となっているヒューリスティックに基づいたルールについては、確率モデルに基づくルールとの比較検討が必要であるが、現実のシステムで性能を向上させようとすると、ヒューリスティックによる方法になることが多く、実用上有効な技術であると思われる。

[参考文献]

[Nagata 1994] Nagata :A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, COLING94 Proceedings Vol.1, pp. 201-207 (1994)

[小松 1990] 小松, 安原 ：コスト最小法形態素解析のコストルールの作成方法, 自然言語処理研究会資料, 85-1, pp. 1-8 (1990).

[久光 1990] 久光, 新田：接続コスト最小法による形態素解析の提案と計算量の評価について,電子通信学会技術報告 NLC90-8 pp. 17-24 (1990).

[久光 1991] 久光, 新田：接続コスト最小法による形態素解析,情報処理学会第41回全国大会予稿集 Vol.3 pp. 1-2 (1991).

[丸山 1994] 丸山, 荻野 ：正規文法に基づく日本語形態素解析, 情報処理学会論文誌 Vol.35 No.7, pp. 1293-1299 (1994).

[吉村 1989] 吉村, 武内, 津田, 首藤 ：未登録語を含む日本語文の形態素解析, 情報処理学会論文誌 Vol.30 No.3, pp. 294-301 (1989).

付録1：第1グループのトレーニング・データ（131文）

「地震、雷、火事、おやじ」という言葉。10月6日。10兆円台。12人。1人。ODA予算の伸び率。あみあげ型式。ある登山に関する本。おへそ。この決議。この言葉。こわさ。ひもつきの借款。もの。わが国。わが国のODA。わが国の援助。わが国の輸出。アフリカの後発途上国。ゴルフ場での事故。サーフィン中の高校生ら。サッカー部員11人。フィリピン・レイテ島の農家。マメ。マルコス疑惑のような不祥事の再発。愛知県の中学校の校庭。意図。一度。円借款事業。援助。援助の原資。援助の分野。援助活動。援助増強。援助地域。夏の暑さ。夏休み。夏山。蚊帳。会の活動。海。海外からの研修生の受け入れ。海辺。活動の輪。関係機関。帰国後の職場復帰。技能。共感。胸。空。靴。啓発活動。経験。月200円の会費。研究。個人の善意。午後。考え方。高知県の海岸。国際協力の日。国際協力事業団。国際世論。国民総生産。今回の事故。今年。山。山の天気。死傷者。私たち。事前の調査。自然との接し方。自然に対する畏敬の気持ち。自然のこわさ。若い隊員。出来事。準備や注意。初心者の心得。助成措置。小屋。小屋の人。人。人の交流。人間味。水牛。世界各地。政府。政府の援助。政府の予算編成。政府開発援助の規模。生活苦。青年海外協力隊。税金や郵便貯金。先輩。増額。大型プロジェクト。知恵。地図。中高年の隊員。中年以上の人。注意。途上国。途上国の自助努力。途上国援助や経済協力に対する国民の関心。読者。日本の対外援助額。日本企業名の公表。農耕用の水牛。発想。避雷の力。米国自身の援助疲れ。米上院。保健活動。母親や年長者。防衛費。防衛費そのもの。防衛費の肩代わり。防衛費の増強。無医村。無償援助の割合。面白さ。問い合わせや拠金。有償の円借款。遊びや楽しみ。幼児体験。要求。来日中の研修生。雷。雷さん。雷雨。落雷事故。例。

付録2：第2グループのトレーニング・データ（101文）

1950年代、。20年も前。24時間体制。3年前、。すぐれた手本。そのかなめ。その要点。ぼけ。やる気と実績のある人びと。わが国の精神医療。アパートや住宅。オーナー。カナダのバンクーバー。カナダやイタリア。カネ。スタッフ。チーム。フランス。モデル地区。悪名。圧力。意図。入院の必要。入院患者。入院者の6割。永久下宿人。英国。家や自殺防止センター。改革。海外。勧告。患者。患者の職業訓練や職探し。気。虐待。後進性。公費。厚生省。厚生省調査による推計。合意。国会。国際人権規約。国際水準。国際的非難。国民の間。国連の小委員会。今。今回の改正のきっかけ。今日。私立病院の民間活力。持ち出し。社会。社会復帰という言葉。社会復帰活動。手だて。手段。収入。小都市や町、村。情緒障害の子どもたち。条文。人。人たちの共同住居。人手。世界の動き。世界保健機関。政策。政府や自治体。生活訓練施設や授産施設。精神衛生法。精神科救急診療のための

家。精神科専門の看護者やソシアルワーカー、作業療法士、臨床心理士。精神科訪問看護婦やソシアルワーカーなどマンパワーの充実。精神病院。精神保健法。西欧諸国。税金。退院者や通院者の心と生活。地域精神保健医療の見本と改革のプログラム。町なか。当時の勧告。日常生活や職業生活。日本。日本の精神医療。日本の精神衛生法。日本の精神病院。日本国。日本政府批判。病院内。分野。閉じ込め型。閉じ込め路線。閉鎖性。方式。法案。訪問看護。本人の自由意思による任意入院。無知。名称。妄想。約３０人ずつ。立場の人たち。

付録３：第３グループのトレーニング・データ（１０４文）

１１地域の公害防止計画の内容。１３年。６日の公害対策会議。９万６７００人。このところ。この結論。この種の公害の先進国アメリカ。この節目。これから。これまでの枠。こんどの地域指定解除。そこ。そのほか有害化学物質などへの対策。その際、。その費用。ひとつの節目。わが国の環境行政。ゼロ。ハイテク汚染、フロンガス、アスベスト、。意味。医療費や障害補償費。一定期間。一歩。煙突型公害の被害救済。河川の水質汚濁など都市型、生活型の公害。各自治体。幹線道路ぞいの窒素酸化物。幹線道路沿いの広域的な交通公害、。患者。環境。環境行政のむなしさ。環境庁。環境破壊の懸念。企業。気管支ぜんそく。規制。逆の方向。結果。健康相談などの事業。健康被害との関係。見直し。個人への被害補償。後手。公害。公害という点。公害の対策と予防。公害企業など。公害健康被害補償制度の再発動。公害健康被害補償法（公健法）の改正。公害行政の予見性。公害対策基本法。公害類型。公健法。公健法の再検討。行政。今回の公害防止計画の策定。今後の調査と常時監視。再開発。最近の２つの公害行政。産業構造。仕組み。姿。自治体。自動車排ガス。首都圏、大阪、名古屋、。除去技術の開発。場面。制度。制度の不合理な面。政府。生活排水による湖沼、。先の国会。全国４１カ所の指定地域。素材型から加工組立型。早手回しの対策。対策。大気汚染。大気汚染の実態。大気汚染被害。大都市の主要幹線道路沿いの汚染。担当者自身。地域。地域住民の健康被害。窒素酸化物。展開。都市における大気汚染、。土地利用。東京、大阪など大都市。東京湾埋め立て。当面の緊急課題。二歩。認定患者。年間１０４４億円。年間約９０００人ずつ。被害者の償り。分野。補償費用。法改正。北九州など１１地域。未然防止の配慮。目。問題。硫黄酸化物。硫黄酸化物など。

付録４：評価用データ（１４４文）

４氏の外交政策。４氏の政策。４人の候補。５年間の中曽根政治。８日。８日の告示後。くつかの注文。この種の政見。これまでの政治。これらの議員の行動。その実現までの手順や財政的な裏付け。その停滞。その辺のところ。それぞれの個性と政治スタイル。それぞれの視点。ふるさと創生。よどみやひずみ。わが国の首相。グローバル・ニューディールの提唱。サラリーマンの定期異動。スローガン。ビジョン。ペルシャ湾の安全航行確保。マイナス面。安倍総務会長、宮沢蔵相、竹下幹事長、二階堂前副総裁の政策。影響。英文の資料。宴会政治の自粛。改革への情熱。外国の関心。外国人。外国人記者の取材。各候補。各候補の政策や政治姿勢。各陣営。各陣営のこれまでの動き。各派の駆け引きの場。各派閥。活気。肝心のところ。及び腰の政権争奪。宮沢氏。金丸副総理。金銭。経済力。憲法問題。現職総裁の失政。個々の議員。個所。候補者間。公職選挙法。行事。国の基本法。国公有地の払い下げ。国際社会。国際的な評価。国際的関心。国際的役割。国民。国民の願い。国民資産倍増計画。今度の総裁選挙。産業構造の転換。指導力。持ち味。時間。時期の総裁選び。自民党の総裁選び。自民党の総裁選挙選挙管理委員会。自民党の派閥。自民党所属国会議員４４５人の過半数。自民党総裁選挙。自民党内。自由な意思に基づく投票。首都圏の地価の狂乱。住宅元年宣言。所要の措置。諸悪の根源。焦点。新しい世代。世界。世界の安定や繁栄。世間の目。政局転換。政権。政権の行方。政権交代の意義。政権担当。政策論議。政治のあり方。政治のビジョン。生活大国の設計。選挙管理委員会。創造的外交。争い。総裁公選規程。総裁選び。総裁選挙。総裁選挙の選挙管理。総裁選挙の本選挙。総裁任期満了による選挙。総理・総裁。多数派工作。第１の注文。第２の注文。地価問題、税制改革など、。中曽根政治。中曽根政治の継承。中曽根内閣。中曽根派、河本派の議員や無派閥の議員。鉄の結束。党内の論理。党内行事。投票の自由意思。東京集中と地方経済の停滞。当事者たちの話し合い。内閣や党のポスト。内向きの政治。内政面。肉声の答え。日程繰り上げ問題。日米関係。日本。日本のかじ取り。日本の首相の言動。日本の進路。派閥のお祭り。派閥の決定。派閥の合従連衡。派閥の対応。派閥や個人間での金銭授受の禁止。派閥幹部。派閥間の密室の取引。迫力。不祥事。部分。文化経済国家の創造。平和協力外交。米ソ首脳会談。保守政治の活性化。無記名投票。役割。立候補。老後のためのシルバールネサンス構想。