別紙 1

論文の内容の要旨

論文題目	反辞書の効率的な構築手法に関する方法
学位申請者	深江　裕忠

情報技術の発展により，近年では様々な機器にセンサーが搭載され，センサーで検知したデータに従って動作する機器が多くなってきた．センサーは，温度，振動，超音波，電波など様々なものがあるが，取得するデータは数値の系列である．この数値の系列（入力系列）に対して，検出や解析を行うために，入力系列に出現する全ての部分系列を登録したデータベース(ここでは辞書と呼ぶ)がよく使われる．

本論文で研究対象にしている反辞書とは，従来の一般的な辞書とは異なり，入力系列に出現しない系列（極小禁止語）を登録する辞書である．極小禁止語とは，それ自身は入力系列に出現しないが，極小禁止語の真の部分系列は全て入力系列に出現する系列である．すなわち，入力系列に出現しない系列の中で極小性をもつ．反辞書は，全ての部分系列を登録する辞書よりも，登録する系列が少ないという性質をもつ．また，反辞書を構成するデータ構造として，極小禁止語を受理するオートマトンや，接尾辞木を拡張して極小禁止語のノードを追加するなど効率的な検索方法が提案されている．

反辞書は，Crochemoreらが2000年に提案した無歪みデータ圧縮法に初めて用いられ，符号化で使用される確率モデルの作成において，その有効性が示された．データ圧縮以外にも，反辞書は，分岐予測，同期符号，不整脈検出などへ応用されている．

しかし，反辞書を構築するには，計算量が入力系列長に比例することが知られているが，実用上は，膨大な記憶量と計算時間を必要とする．例えば，太田・森田によって提案された接尾辞木を用いた反辞書構築手法では，アルファベットサイズが多値の場合，記憶量が膨大になる．計算機実験では，アルファベットサイズが256の場合，入力系列長の約1万倍の記憶量を必要とした．

そこで，本論文では，記憶量と計算時間を改善する新しい反辞書構築手法を二つ提案する．

一つ目は，入力系列を全て入力後に反辞書を構築する静的な手法である．二つ

目は，入力系列を逐次的に読み込みながら，反辞書を変化させていく動的な手法である．構築手法を考える場合，一般的には，動的な手法よりも静的な手法の方が簡単である．そこで，最初に静的な手法について検討を行い，その成果を動的な手法へ活用する．

静的な手法では，新しい反辞書の別表現を提案する．この別表現では，極小禁止語を，先頭記号と後続する系列に分けて考えることにより，あと1記号先頭に加えることで極小禁止語になる系列が，入力系列の接尾辞配列ならびに高さ配列（本論文ではL配列と呼ぶ）によって特徴づけられること，すなわち，先頭に１記号加えれば極小禁止語になる系列は，接尾辞配列上の接尾辞の先頭部分に存在し，その長さはL配列で求めることができることを示す．ここでL配列とは，接尾辞配列上で隣接した二つの接尾辞に共通する最長接頭辞の長さの配列である．

さらに，本論文では，被覆集合を提案する．被覆集合とは，入力系列の任意の部分系列が，接尾辞配列上の接尾辞の先頭部分に存在している範囲を表現する集合である．接尾辞配列上の接尾辞は辞書式順序昇順で整列しているので，被覆集合の範囲を比較することで，入力系列の部分系列同士の辞書式順序昇順の大小関係を判別できる．すなわち，この2つの系列の先頭から順に記号を比較する必要がなく，定数時間で系列の比較が行えるようになる．

この反辞書の別表現と被覆集合を用いることで，接尾辞配列，L配列をそれぞれ3回走査するだけで，先頭に１記号加えれば極小禁止語になる系列をすべて含む高々$2n+2$個（nは入力系列長）の系列を辞書式順序昇順で求められる算法を提案する．提案方法においては，極小禁止語は，利用した被覆集合の範囲をそのまま活用して，被覆集合の範囲と系列長，そして，先頭に加える記号を要素に持つ配列で表現される．本論文では，この極小禁止語の配列表現を反辞書配列と呼んでいる．

そして，計算機実験によって提案した算法の有効性を確かめたところ，接尾辞木を用いた従来手法と比較して，計算時間が約20分の1，記憶量が約2.5分の1に改善されることが分かった．

動的な手法では，反辞書そのものではなく，反辞書オートマトン，すなわち，反辞書に含まれる極小禁止語を受理するオートマトンを，入力系列を逐次的に読み込みながら動的に構築する手法を提案する．

まず，すでに読み込まれた入力系列の末尾に新たな記号が加わることにより，反辞書がどのように更新されるかを明らかにした．申請者が修士論文で得た知見である，反辞書の更新において，ある特定の極小禁止語が一つ削除されることと，新たに加わる極小禁止語（新規極小禁止語）は，その両端のどちらか一方に削除される極小禁止語を含む，という二つの性質に加え，本論文では，新規極小禁止語の長さを削除される極小禁止語の長さで上と下から評価した．

つぎに，反辞書の別表現と同様に，あと1記号加えることで新規極小禁止語になる系列を，入力系列の末尾から探索する．さらに，更新前の反辞書オートマトンを用いて，探索した系列を新規極小禁止語にするために加える記号を求める．これらの結果を用いることによって，反辞書オートマトンの構築するために，従来の手法では，接尾辞を表す木構造やオートマトンを経由して作成する必要があったのが，その手間を省いて，入力系列から直接構築する手法を提案する．

別紙２

論文審査の結果の要旨

学位申請者氏名　深江　裕忠

審査委員主査　森田　啓義

委員　長岡　浩司

委員　多田　好克

委員　笠井　裕之

委員　古賀　久志

本論文は，入力データ列に出現しない極小禁止語からなる反辞書の効率的な構築法として，静的な構成法と動的な構成法のそれぞれにおいて新たな手法を提案し，その有効性を理論的ならびに計算機実験によって示している．

本論文は全６章から構成されている．

第１章は，序章であり，研究背景と目的，本論文の主要な結果について述べている．

第２章は，用語の定義や基本的なデータ構造について述べている．とくに，極小禁止語，反辞書の厳密な定義に加え，提案法で用いられている既存の接尾辞配列やＬ配列（高さ配列ともいう）の紹介，ならびに，本研究で新たに導入された被覆集合の概念が説明されている．

第３章は，反辞書構築に関する従来法を紹介している．とくに，本論文の提案手法と密接に関連する接尾辞木を用いた静的な反辞書構築手法について，詳しく論じている．

第４章は，本論文の主要結果の一つである，入力データ系列を全て読み込んだ後に反辞書を構築する静的な構成法について論じている．

４．１節は，前章で紹介した接尾辞木を用いた反辞書構築法では記憶量の大半がＭＦ－linkと呼ばれるポインタが占めていることを述べている．

４．２節は，接尾辞木よりも少ない記憶量で接尾辞を表現できる接尾辞配列を用いるために，極小禁止語を先頭記号と後続する系列に分けて考えた，反辞書の別表現を与えている．

４．３節は，極小禁止語の先頭文字を削除して得られる系列をすべて含む，ある系列の集合（ここではＧと呼ぶ）が，接尾辞配列とＬ配列を用いて求めることができることを示している．

４．４節は，前節までの結果を用いて，提案する静的な反辞書構築法の概要を述べている．

４．５節は，集合Ｇに含まれる系列の被覆集合について，Ｌ配列を用いて被覆集合に属する系列の長さを評価している．

４．６節と４．７節は，接尾辞配列とＬ配列から集合Ｇを構築する詳しい手順を与えている．

４．８節は，提案法の計算時間が入力データ系列の長さに比例することを証明している．

４．９節は，提案法の記憶量が入力データ系列の長さに比例することを証明している．

４．１０節は，提案手法と接尾辞木を用いた従来手法とを計算機実験で比較している．実験結果から，提案手法は従来手法に比べ，反辞書構成のためのコストは計算量で約1/20，記憶量で約2/5に改善されることが示されている．

第５章は，入力系列を逐次的に読み込みながら，反辞書を変化させていく動的な手法を論じている．

５．１節は，本章で論じる動的な反辞書構築手法が，接尾辞木や接尾辞配列などの中間データ構造を介さず，入力データ系列から直接，極小禁止語を受理する反辞書オートマトンを構築することを述べている．

５．２節は，本章で新たに用いる表記を追加している．

５．３節は，反辞書オートマトンについて例を用いて説明している．

５．４節は，申請者が修士論文でもとめた，動的な反辞書構成に関する既知の結果を述べている．

５．５節は，入力データ系列の長さが新たに１記号分だけ増えた場合に，新たに反辞書に加わる極小禁止語の長さの上限と下限を評価している．

５．６節は，第４章の反辞書の別表現と前節までの考察に基づき，入力データ系列に新しい記号が加わったときに，反辞書オートマトンを更新させる手順を示している．

第６章は本論文で得られた成果を総括し，今後の検討すべき課題について述べている．

以上をまとめると，本論文は，静的な場合と動的な場合の両方において，反辞書の新しい構成法を示している点，理論と計算機実験の両面から提案手法の性能を評価した点において優れており，今後，これらの構成法はデータ圧縮や情報検索など関連分野においてさまざまな応用が期待される．よって本論文は博士（工学）の学位請求論文として十分な価値を有するものと認める．