決定性有限オートマトンと文字候補ラティスの照合によるフレキシブルOCR知識処理

Flexible OCR Postprocessing by Matching Deterministic Finite Automaton and Character Candidate Lattice

鈴木　智久　　中島　康裕
■SUZUKI Tomohisa　■NAKAJIMA Yasuhiro

OCR（光学的文字認識）の高精度化には認識対象（姓名や，住所，電話番号など）に固有の知識処理が有効である。多くの場合，知識処理としてはリストに登録された単語と合致するように認識結果を補正する単語後処理が行われるが，近年は単語後処理では対応できない部分を含む複雑な認識対象の読取りのニーズが高まっている。従来，このような複雑な認識対象を拡充する際には認識対象ごとに知識処理のプログラムの開発や改造が必要であった。そこで東芝は，認識対象を記述した正規文法を元に知識辞書を生成することで，認識プログラムを変更することなく容易に認識対象を拡充できる技術を開発した。

An effective means of improving the accuracy of optical character recognition (OCR) is to apply postprocessing using specific knowledge of the recognition targets, including full names, addresses, and telephone numbers. Word postprocessing is generally performed by matching the recognition results with words previously registered in a dictionary. Demand has been growing in recent years for the application of word postprocessing to more complicated recognition targets. However, it is difficult to reduce the costs and shorten the development period required for the modification of post-processing programs in order to handle such complicated word matching.

With this as a background, Toshiba has developed a flexible OCR postprocessing technology that implements word postprocessing by matching a deterministic finite automaton (DFA) converted from a regular grammar describing the recognition target and a candidate character lattice. This technology makes it possible to easily expand recognition targets without changing the program, and is currently being introduced into OCR products for effective processing of various finance- and insurance-related documents.

1　まえがき

OCRでは通常，光学スキャナで読み取った紙面の画像から文字やその候補を切り出して，認識辞書に収めたパターンとその各々とを照合することで認識する（**図1**）。個々の文字の認識では，ノイズや画像劣化の影響による誤読のほか，文字単体を見ても区別がつかない同型文字や類似文字との誤読も生じるため，認識対象（姓名や，郵便番号，住所，電話番号など）として意味が通るように誤りを補正する“知識処理”による高精度化が行われるのが一般的である。更に，手書き文字やプロポーショナルフォントによる印字を認識するときには，認識対象の背景知識なしには個々の文字の位置すら特定できない場合があり（**図2**），知識処理がより重要となってくる。

この知識処理の手法としては，あらかじめ作成した単語リスト中の単語のいずれかと合致するように認識結果を補正する“単語後処理”をベースとした手法が数多く提案されており，効果を上げてきた。

しかし近年，数量など単語後処理では対応できない部分を含む複雑な認識対象の読取りのニーズが高まっており，認識対象に応じた専用プログラムの開発に掛かるコストや期間が問題となっている。

そこで今回，専用プログラムの開発を行うことなく，複雑な認識対象の知識処理を実装できる，フレキシブルOCR知識処理技術を開発した。開発した技術では，認識対象の文字の並びを記述した正規文法から知識辞書と呼ばれるデータファイルを生成し，認識プログラムをこの知識辞書で駆動することで，認識対象に固有の知識処理を行う。この技術により，ユーザーは正規文法を記述するだけで，認識対象に固有の知識処理を手軽に実装することができるようになったうえ，正規文法で記述できる範囲内であれば，従来の技術では対応できなかった，より複雑な認識対象の知識処理を実現できるようになった。

ここでは，従来の単語後処理とその問題点，及び開発した

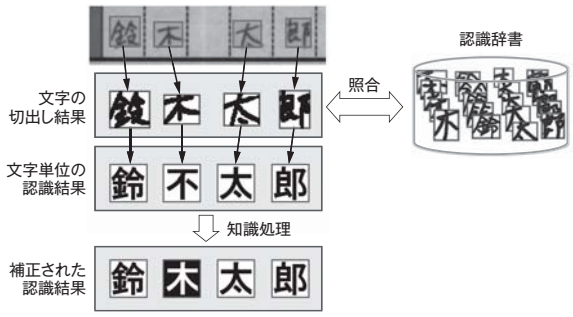

図1．文字認識における一般的な処理の流れ — 切り出した文字を認識辞書中のパターンと照合することで認識する。
Typical flow of OCR

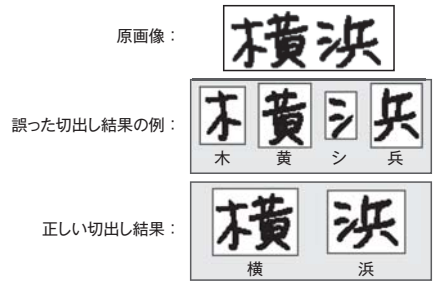

図2．認識対象の背景知識なしには文字の切出しが困難な場合の例 — 誤った切出し結果でも，切り出した文字はそれぞれ文字らしく見える。

Case in which character segmentation is difficult without background knowledge of recognition target

技術の概要について述べる。

2　従来の単語後処理とその問題点

単語後処理を行う認識系の構成の例を**図3**に示す。この構成ではまず，紙面の画像から文字のパーツを検出する。次に，それらのパーツを組み合わせてできる“候補文字”の接続関係を示した“候補文字ラティス”を構成したうえで，ラティス上の候補文字の各々を認識辞書中のパターンと照合することで読みの候補を列挙する。そして，ラティス上で文字列の始点から終点への経路探索を行い，候補文字の読みの候補を1個ずつ選んで並べた結果が単語リストに登録された単語と合致する経路を探す。探索の結果，合致する経路が見つかった場合，その単語を認識結果とする。

この方法による単純な単語後処理は，認識対象として出現しうる文字列を網羅した単語リストを作成するだけで様々な認識対象に適用できる反面，単語リストで網羅できない認識対象には適用できない。

例えば，ひとつづきに記入された姓名（**図4**(a)）においては，わが国ではそれぞれ数万通りもある姓と名の組合せは数十億通りとなり，このように膨大な数の組合せを全て網羅するのは不適切である。したがって姓名の知識処理は，姓と名の区切り位置を推定しつつ区切り位置の前後で別々に行うのが，処理速度及び記憶容量の面で合理的である。

また，住所の認識に単語後処理を適用する場合，一見する

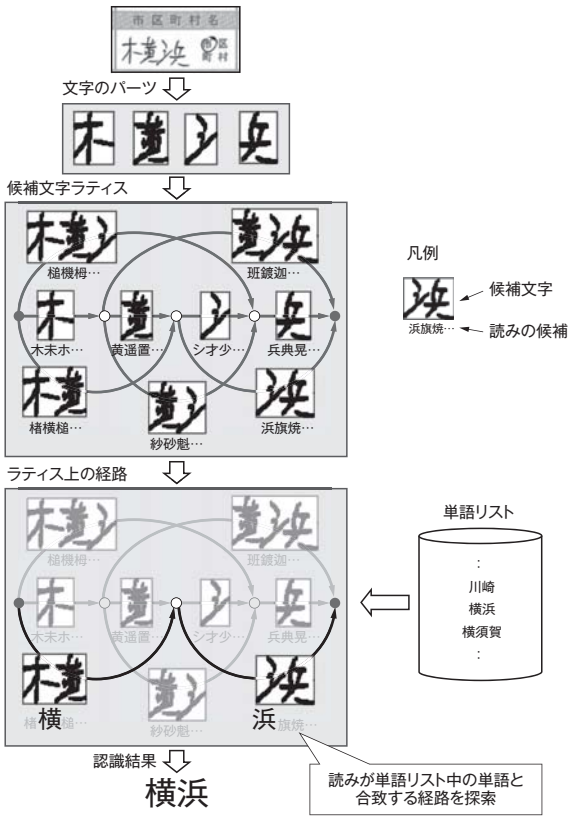

図3．単語後処理を行う認識系の構成例 — 候補文字ラティスを構成して，候補文字の読みが単語リスト中の単語と合致する経路をラティス上で探索する。

Example of typical recognition system with word postprocessing by matching candidate character lattice

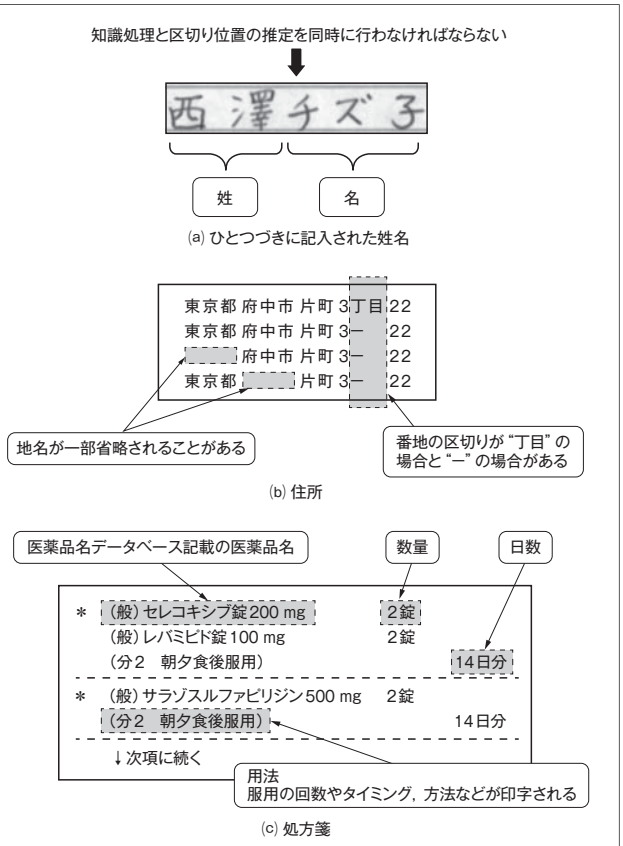

図4．単語リストで網羅するのが不可能又は不適切な認識対象の例 — 複数の単語の組合せや，表記ゆれ，数量などが見られる記入対象は単一の単語リストでは網羅できない。

Examples of recognition targets for which coverage by single word list is impossible or inappropriate

一般論文

と全国の住所を網羅したリストを作成すればよいように思われるが，実際の住所記入では地名が一部省略されたり，“丁目”や“番地”などの番地区切りが“−”と略記されたりと，表記ゆれが見られるため，表記ゆれの組合せを全て網羅した単語リストを作成するのは困難である（図4(b)）。

更に知識処理が困難な認識対象として挙げられるのが，処方箋の記載内容である。処方箋では，医薬品データベース記載の医薬品名や，用法，用量，備考などの記載項目が混在して印字されている（図4(c)）。これらの項目は一定の構文に従って印字されていることが多いものの，項目ごとの印字位置が一定しておらず，単位や用語の表記ゆれも見られるため，項目ごとに印字位置を精度よく推定したうえで項目別に単語後処理を行うのは非常に困難である。したがって処方箋を認識する場合には，項目別に単語後処理を行うのではなく，医薬品名などの単語のほかに，数量や用法，備考など単語リストでは網羅できない部分も併せて，構文に従った知識処理を行うのが好ましい。

このように単一の単語リストでは網羅できない認識対象の知識処理の従来手法としては，文字列の先頭から末尾まで順に単語後処理を行っていく方法[1]や，あらかじめ単語を検出しておき，単語の並び方の規則に従ってそれらの単語を連結する方法[2]-[4]，単語リストと合致した箇所で単語後処理を行い，合致しない未知語領域においてn-gram後処理（自然言語における文字の並びの統計的な傾向に基づいた後処理）を行う島駆動型文字列認識方式[5]などが知られている。

共通のアルゴリズムで様々な認識対象の知識処理を統一的に実現できると，認識対象の拡充のたびにプログラムを修正する必要がなく好つごうである。そのため，前述の順に単語処理を行っていく方法や単語を連結する方法には，認識対象に合わせてカスタマイズできる文脈自由文法で単語の並び方の規則，すなわち構文を規定するようになっている手法[1], [3], [4]がある。しかし，それらの手法は認識対象が単語を連結したものとなっているという前提に基づいているため，単語リストでは網羅できない部分の知識処理を行うことができなかった。

それに対して，島駆動型文字列認識方式では単語リストでは網羅できない部分の知識処理を行うことができるものの，構文に従った知識処理を行うことができなかった。

また，文脈自由文法で認識対象を規定する手法[1], [3], [4]では，知識処理の実行時に文脈自由文法を構成する書換えルールを展開する必要があり，このルールの展開が計算量的な負荷となっていた。

このように従来の手法では，単語リストで網羅できる部分と，数量など単語リストでは網羅できない部分が混在した複雑な認識対象に対して構文に従った知識処理を効率的に行うことができなかった。

3　正規文法による知識処理

開発した技術では，プログラムを変更することなく認識対象へのカスタマイズが行えるよう，正規文法で認識対象の文字の並びを規定し，それを候補文字ラティスと照合することで知識処理を行うことにした。

認識対象を記述した正規文法の例を**図5**に示す。この例では，数字1桁や，医薬品名，項目番号など様々な認識対象の文字の並びのルールを名前付きで定義したうえで，それらを参照する形で処方箋の読取り行など，より複雑な認識対象のルールを定義している。また，医薬品名など単語リストで記述できる部分については，単語リストを収めたテキストファイルへの参照を記述するようになっている。このように，この正規文法はOCRの専門知識を持たない人でも，認識対象の知識を駆使して知識処理を柔軟に実現できるようになっている。

この技術による知識処理結果の例を**図6**に示す。この技術では，単語だけではなく文字の並びも正規文法で規定するようになっているため，単語リストでは網羅できない番号や数量なども取り扱うことができ，番地付きの住所や処方箋のように単語リストで網羅できない部分を含む複雑な認識対象でも，構文に従った知識処理を行うことができる。

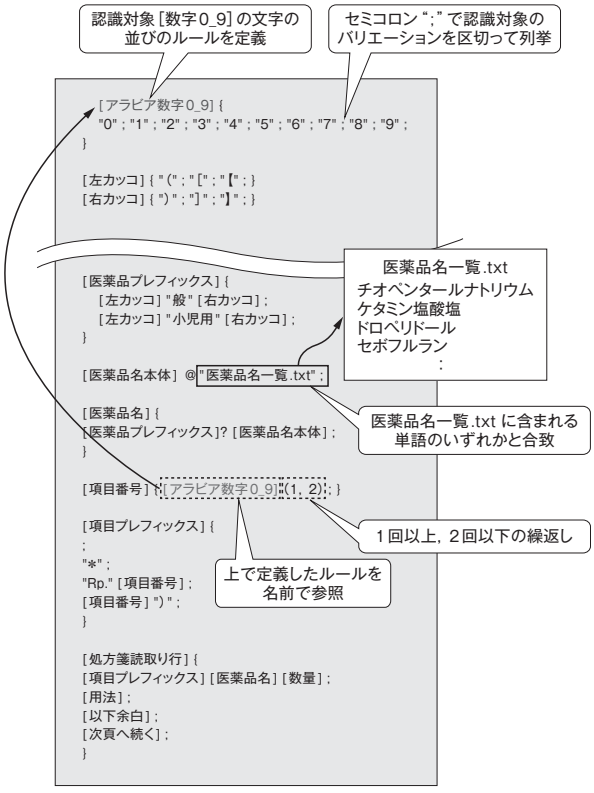

図5．認識対象を記述した正規文法の例 — 認識対象の知識を反映した正規文法を記述することで，知識処理を実現できる。

Example of regular grammar describing recognition target

図6. 知識処理の結果の例 — 単語リストで網羅できない番地や数量などでも知識処理で誤りを補正できている。

Examples of results obtained by postprocessing

4 正規文法の決定性有限オートマトンへの変換

開発した技術では，認識対象の文字の並びを記述した正規文法を決定性有限オートマトン（DFA：Deterministic Finite Automaton）に変換することで，知識辞書を生成する。そして，知識辞書中のDFAを候補文字ラティスと照合することで，認識結果の文字の並びが正規文法で記述したとおりになるように知識処理を行う。開発した技術における知識処理のようすを**図7**に示す。

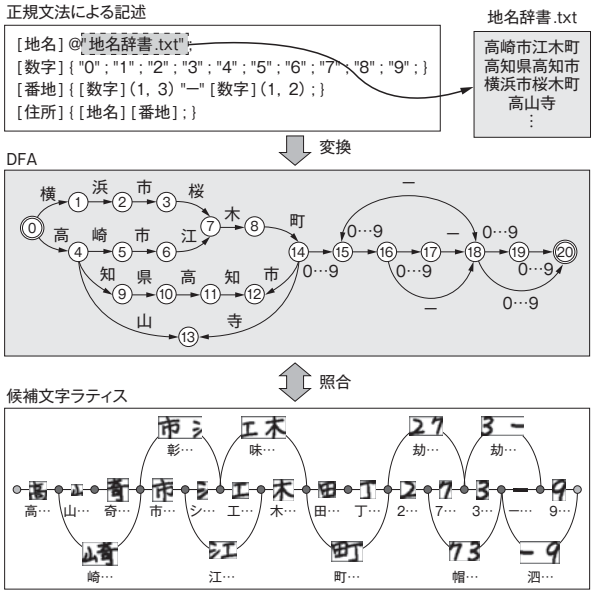

図7. 開発した技術の概念 — DFAと候補文字ラティスを照合することで，DFAの元の正規文法と合致する認識結果を得る。

Outline of flexible OCR postprocessing

この技術では，文脈自由文法で認識対象を規定する既存の手法[1), 3), 4)]で知識処理の実行時に行っていたルールの展開を，正規文法のDFAへの変換であらかじめ行っておくことで，知識処理のプログラムの負荷を低減し，高速化を図っている。

また，DFAは状態を表すノードと，それらの間での遷移を表すエッジだけで表現されており，変換元の正規文法に見られる単語やルールの境界が消失したシームレスな構造となっている。したがって開発した技術では，それらの境界をまたいで知識処理を行っているにも関わらず，境界に関わる特別な処理をいっさい行う必要がなく，簡素で統一的なアルゴリズムを適用できた。

5 あとがき

正規文法で認識対象を記述するだけで，そこから知識辞書を作成し，プログラムの開発を伴うことなく知識処理の対象を拡充できる技術を開発した。開発した技術によって，従来は知識処理を行うことが難しかった複雑な認識対象でも，知識処理を体系的な記述で容易に実装できるようになった。

この技術は東芝ソリューション（株）のOCR製品に搭載されており，金融や保険業界で取り扱われている各種書類の認識で効果を上げている。現在，読取対象の拡大を行っており，薬品名や型式などの読取りの実用化が期待されている。

今後は，正規文法で記述した規則に加えて，n-gram後処理の統合や，不適切な認識結果を禁止パターンで抑制する技術の実装を行う予定である。

文 献

(1) 寺崎正則 他．“自由記載姓名文字列に対する知識処理”．情報処理学会 第41回（平成2年後期）全国大会講演論文集(2)．仙台，1990-09，情報処理学会．1990，p.2-208－2-209.

(2) 丸川勝美 他．手書き漢字住所認識のためのエラー修正アルゴリズム．情報処理学会論文誌．**35**，6，1994，p.1101－1110.

(3) 橋本雅美 他．“構成文法を用いた複合語の文字認識後処理方式”．電子通信学会 情報・システム部門全国大会．1985-11，電子通信学会．1985，1-68.

(4) 高橋寿一 他．回帰的遷移ネットワークを用いた文字経路探索方式の開発．電子情報通信学会技術研究報告 PRMU．**109**，418，2010，p.141－146.

(5) 嶺 竜治 他．N-gram言語統計量を併用した島駆動型文字列認識方式．電子情報通信学会論文誌．**J89-D**，5，2006，p.1011－1018.

鈴木 智久 SUZUKI Tomohisa

インダストリアルICTソリューション社 IoTテクノロジーセンター 知識・メディア処理技術開発部主務。文字・画像認識の研究開発に従事。電子情報通信学会会員。

IoT Technology Center

中島 康裕 NAKAJIMA Yasuhiro

東芝ソリューション（株）プラットフォームセンター ハードウェア開発部主任。OCRソフトウェアの製品開発に従事。電子情報通信学会会員。

Toshiba Solutions Corp.

一般論文