和漢書用図書カード認識システム

5W-4

松田　充弘　　米田　政明　　長谷　博行　　酒井　充

富山大学工学部電子工学科

1．はじめに

図書館の電算化に伴い、図書目録カード（以降、カード）の遡及入力が行なわれている。しかしながら現状ではその入力を人手に頼らなければならない。またその量も莫大なものである。

本研究室では、この問題に対し洋書用図書カードの自動認識システム（以降、前システム）を作成したが(1)、これを発展させた和漢書用図書カード認識システム（以降、新システム）を現在試作している。そのシステムの考え方について述べる。

2．図書目録カード

カードは、目録規則に従って記載されている。一般に、洋書用カードは英米目録規則(2)、和漢書用カードは日本目録規則(3)に基づいている。これらの規則に定められた各項目は、順序が定められていて、かつ省略可能である。また項目間の区切り記号も定められていてこれにより項目が分離されている。しかし、洋書用カードでは区切り記号が「:」「;」「,」等の特殊文字であるのに対して、和漢書用ではスペースのみが区切り記号である。さらにスペースは区切り以外にも使用されるので、カードを理解するときには意味内容も合せて考えなければならない。

両システムとも、富山大学付属図書館で使用されているカードを対象にする。図書カード例を図１に示す。

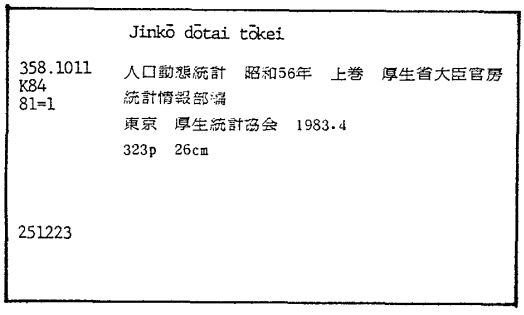

図１．和漢書用図書カードの例

3．前システムとその問題点

3．1　前システムの概要

前システムは、英文タイプライターで記載された洋書用カードの認識のために作成された。このシステムは、ＦＯＲＴＲＡＮ言語のみで書かれている。システムは、前処理・文字認識・項目分類の３つの段階から構成され、これらを順に実行する（図２）。

前処理部では、１次矩形抽出・２次矩形抽出・行分離・３次矩形抽出やノイズ消去を行ない、後の文字認識・項目分類で使用する矩形の位置と順序情報を生成する。

文字認識部では、文字の切出しと認識をＤＰを用い、接触・分断等を考慮に入れて行なう。

項目分類部では、各項目ごとに項目規則を満たす部分文字列をすべて求め、項目の順序と文字列の順序を利用し、非決定的に最適な項目分類結果を求める。

3．2　前システムの問題点

前システムの問題点としては、

１）システムの動作が完全自動でオペレータの介在がないため、ノイズ等で品質の良くないカードでは認識できない場合がある。

２）項目規則がプログラム中に書かれているため、変更するのが困難である。項目規則は洋書用と和漢書用では当然異なり、また年代によっても、図書館によっても細部が異なっている。

３）英文用タイプライターで記述されたカードのみを対象にしているので、手書きや漢字で記述されたカードには対応できない。

などの点がある。

ただし問題点１に関して、ポインティングデバイスを使用したカードのノイズ除去や項目の位置指定については、本研究室で実験を行ない良好な結果を得ている。

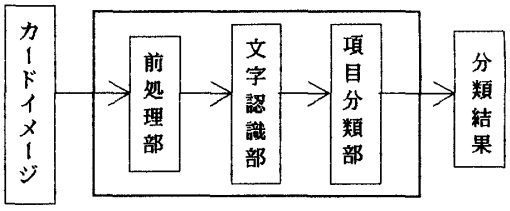

図２．前システムの概略図

A BOOK CARDS RECOGNITION SYSTEM FOR JAPANESE BOOKS.
M. MATSUDA, M. YONEDA, H. HASE, M. SAKAI
TOYAMA UNIVERSITY

４．新システムについて

新システムは和漢書用カードも対象にするために、必然的に不確実な項目規則を適用することになり、処理も複雑になるので、前システムのようにＦＯＲＴＲＡＮのみで記述することが困難になる。

４．１　新システムの改良点

まず、洋書用のシステムを和漢書用にも対応できるように変更するために和文タイプライターの文字をサンプルにして、２１７２文字種（記号を含む）の認識を可能にする。

さらに、概念を階層的に記述し項目規則を宣言的に表現するためにＰｒｏｌｏｇ言語を使用する。ただし画像処理等の数値計算も多く要求されるので、この目的のために本研究室で作成した拡張Ｐｒｏｌｏｇ処理系(4)（ＦＯＲＴＲＡＮサブルーチンを容易に呼び出すことができる）を使用する。

４．２　フレームによる表現

カードを画像の面で捉えると、１次矩形（文字単位）、２次矩形（語単位）、３次矩形（ブロック単位）、カード画像（３次矩形群と後述の領域との対応）の様に階層的に考えることができる。またカードを記号処理の面で捉えると、文字、項目、幾つかの項目を集めた領域（書誌内容、分類コード、受入番号の３つの領域がある）、そして図書カードの様にやはり階層的に考えることができる。

これらの概念をフレームを導入してＰｒｏｌｏｇで記述する。フレーム構造は２つの平行した世界をもち、それぞれ画像処理と記号処理に対応する。各々の世界でフレームは上下関係をもち、かつ両世界間は矩形と領域の対応という点で関係をもつ（図３）。

一般的に処理の開始は、図書カードのフレームにインスタンス要求を出すことである。インスタンスが既に生成されていればそれを返す。生成されていなければ、画像世界のカード画像フレームのインスタンスと上述の３領域との対応付けを仮定し（領域仮説）、仮説を検証する。インスタンスが得られたならば仮説の検証に成功したことになるが、もし仮説の検証に失敗した場合は別の領域仮説をたて再度検証する。図書カードフレームからインスタンスを要求された各領域フレームでは、もしそのインスタンスが作られていなければ、下位の項目フレームにインスタンス要求を出す。各項目フレームでは、項目規則に適する文字列をインスタンスとし、上位の領域フレームにその是非を問う。文字列とのパターンマッチングでは誤認識しやすい文字種に対しては仮定付受理(1)を行なっている。領域フレームでは、下位の項目フレームのインスタンスが最も適当なものになるように、別の候補を求めるため項目フレームに再度インスタンス要求を出す。このようにして、階層的仮説すなわち領域仮説、項目仮説、文字の仮説を用いることにより柔軟な処理が実現できるものと考えられる。

４．３　項目規則の表現

項目規則の表現の柔軟さは、そのシステムの柔軟さである。このシステムでは項目規則表現のための新たな言語を定義し、そのインタープリタに条件（規則）を送るとそれとマッチングする部分文字列を返す。この言語の文法には、文字数、キーワード、データーベースなどを含んでいる。キーワードとは、'著'や'編'や'版'などの文字である。また、東京や大阪などの出版地をデーターベース化しておく。条件は項目規則内にＡＮＤ，ＯＲ結合した形で宣言的に記述され、インスタンス要求があったときに引数の形でインタープリタに与えられる。

５．おわりに

本システムは、以前より本研究室で行なわれている図書目録カードの認識の新たな段階である。上述の枠組みは、インタラクティブな環境の実現を容易にし、かつ高速な処理システムが期待できる。さらに実用化に向けて現在のオフライン認識を入力から最終出力までユーザーと対話型で処理できるシステムにしていきたいと考えている。

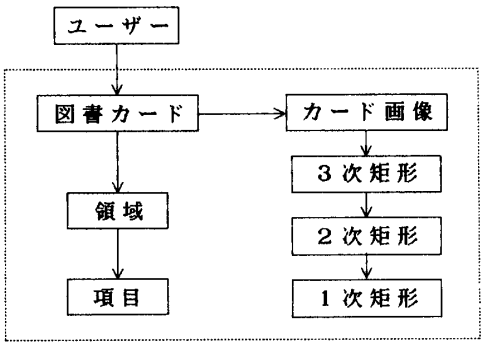

図３．インスタンス要求の流れ

参考文献

(1) 長谷，米田，酒井，吉田；”図書目録カードの自動項目分類について”，信学論，Vol.J70-D, No.8, pp.1579-1588 (昭62).

(2) M.Gorman and P.Winkier編；”英米目録規則”，第２版，日本語版，日本図書館協会（昭57).

(3) 日本図書館協会目録委員会編；”日本目録規則”，日本図書館協会(昭52).

(4) 北野，門村，酒井，長谷，米田，吉田；”ＦＯＲＴＲＡＮと結合可能なＰｒｏｌｏｇ処理系の作成(一)(二)”，電気関係学会北陸支部連合大会，B-8,B-9,pp.97-100 (昭62).