紙面法規文書からＳＧＭＬ文書への変換システムの開発(1)
― 概要と文書認識 ―*

3Ｓ－7

岡本　卓哉　　里　佳史　　村田　英子　　樋野　匡利†
(株) 日立製作所　情報・通信開発本部‡

１．はじめに

筆者らは、紙面法規文書を認識し、構造化文書形式の一つであるＳＧＭＬ(Standard Generalized Markup Language)形式に変換する法規文書入力システムを開発した。本稿では、本システムの機能概要および文書認識処理について報告する。

２．法規文書入力システムの概要

法規文書入力システムは、図１に示すように、

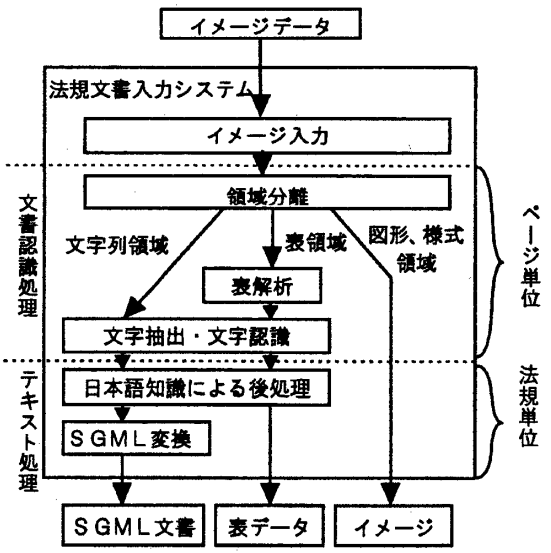

図１　法規文書入力システムの構成

イメージデータとして読み込まれた法規文書に対して、まず、ページ毎に文書認識処理を行い、ページ単位の文字認識結果を得る。領域分離処理では、イメージから図表と文字列の領域を分離する。そして、表解析処理で表領域をセル(表を構成するフィールド）に分割する。文字抽出・文字認識処理では、文字列領域及び表の各セルに対して文字抽出及び文字認識[1]を行なう。

ページ単位に得られた文字認識結果は、例規毎にまとめられ、テキスト処理が行われる。ＳＧＭＬ変換処理では、文字の位置情報及び後処理後の文字認識結果から、文書の構造解析のためのキーとなる文字列を抽出し、この文字列を手掛かりに、構文のチェックを行い、埋め込むべきタグの種類と位置を決める。図と表は、それぞれイメージファイルと表データファイルに変換し、ＳＧＭＬ文書の該当位置にファイル名を記述したタグを埋め込むことで参照できるようにする。

３．文書認識

領域分離と表解析の処理内容について述べる。

（１）領域分離処理

本システムの処理対象である法規文書は、横書きの場合、図２に示したレイアウトで記述されて

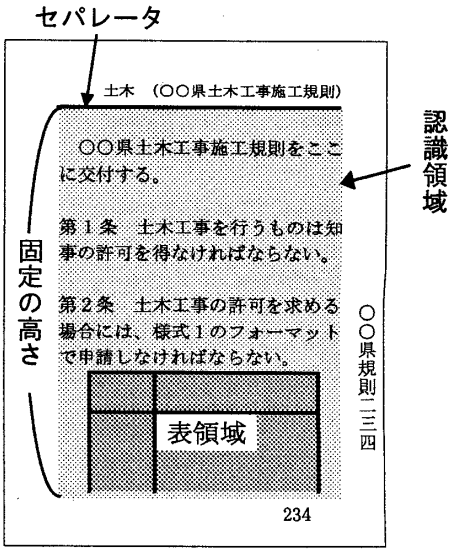

図２　法規文書フォーマット解析

*SGML Conversion System for Regulation Document(1). - Summary and Document Recognition -
†Takuya OKAMOTO , Yoshifumi SATO , Eiko MURATA , Masatoshi HINO.
‡Information Systems R&D Division,Hitachi,Ltd.

いる。ページ内に記述されている情報のうち、ヘッダやページは、本文と切り分ける必要がある。これらの部分をあらかじめ除去することで、領域分離の処理を容易にする。

まず、ページ上部の本文とヘッダを分けるセパレータを抽出する。セパレータを上辺として、あらかじめ定められた固定の高さを持つ矩形領域を設定し、この領域を認識領域とする。

この領域内で、黒画素の連結成分を抽出し、あらかじめ設定された閾値より大きな連結成分の領域は、図表領域とする。さらに図表領域の内部に罫線が含まれるか否かを検出することで、図と表を判別する。

（2）表解析処理

表領域から、罫線を抽出し、抽出された罫線で表領域を分割することで、表をセルに分解する。罫線抽出は、縦罫線、横罫線の順に処理を行う。縦罫線抽出処理は、図3に示すように、まず、閾

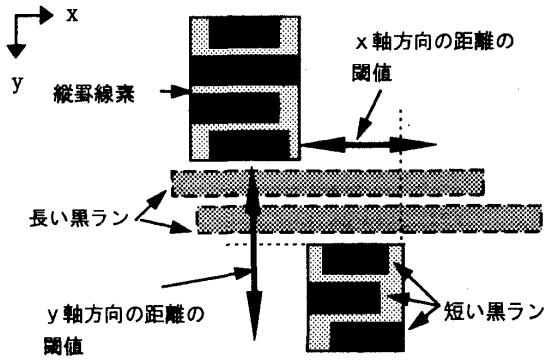

図3　罫線素の抽出

値以下の長さの横方向の黒ラン（連続した黒画素の列）を抽出する。次に、これらの黒ランのうち縦方向に接して並んでいる黒ランを抽出して統合し、縦罫線素とする。さらに罫線の途切れに対応するため、間隔が閾値以下の罫線素を結合する。横罫線も同様の処理を行うことで罫線が得られる。

以上の処理で抽出した罫線のうち、罫線の始点と終点の両端が、表領域の枠あるいは他の罫線と接していれば、表を構成する罫線と判断し、その他は文字の一部であるとみなす。

次に、抽出された罫線によって表領域をセルに

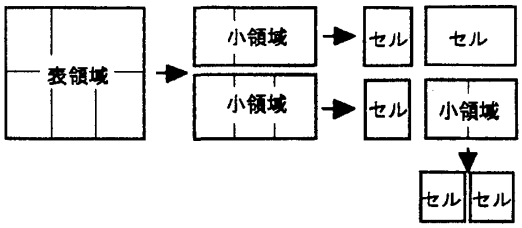

図4　表領域のセルへの分割

分割する処理を行う。図4に示すように、まず、処理領域の初期状態を表領域全体として、領域を分割しうる罫線を探索する。該当する罫線が見つかれば、領域をこの罫線で2つの小領域に分割する。見つからなければ、その領域は分割が終了したものとみなし、セルとして登録する。小領域に対して再帰的に上記の分割処理を行なうことで、表をセルに分割する。

4．評価実験

表解析の評価実験結果を表1にまとめる。本評価は、サンプルの法規文書60ページに対して行なった。解析誤りとなったものは、(1)途切れにより罫線を抽出できなかったもの、(2)セル間隔が狭いため、文字を罫線と誤ったものの2種類である。領域抽出については、全て正しく処理できた。

表1　表解析結果

正しく解析した表	57(95%)
罫線の途切れによる誤り	2(3%)
文字を罫線と判定したことによる誤り	1(2%)

5．まとめ

文書認識を利用した法規文書のＳＧＭＬ変換システムを開発し、その実用上での有効性を確認した。今後は、サンプルデータによるシステムを試用し課題の洗い出しを行い、システムの改良を進めていく予定である。

6．参考文献

[1]「黒画素方向性特徴のずらしマッチングによる印刷文字認識方式の開発」，岡本他，情報処理学会第45回（平成4年後期）全国大会