紙面法規文書からSGML文書への変換システムの開発(2)

3S－8

―文字認識結果のSGML変換―

里 佳史　岡本 卓哉　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、紙面法規文書の文字認識を行い、構造化文書形式の一つであるSGML[1]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、そのサブシステムとして開発した、文字認識結果をSGML文書に変換するSGML変換システムについて述べる。本システムは、文字認識結果から論理構造を表す文字列を抽出して、その文字列を手がかりとした論理構造認識によって文書構造を抽出し、SGML文書を生成するものであり、論理構造認識に用いるルールを対象文書に設定された論理構造定義から半自動的に生成することを特長とする。

2　既存の構造化文書生成技術の問題点とその解決方法

構造化文書形式を採用した文書処理システムに対するニーズが高まるにしたがい、紙面文書やプレーンテキストのような、文書構造を明示的に表現する情報を持たない非構造化文書から、構造化文書を生成する方法についても検討が行われてきている。

従来の方法としては、土井ら[1]や山田ら[2]の方法に見られるように、対象とする文書型の分野を限定して、その分野において共通性のある論理構造(以下「共通論理構造」)と論理構造認識ルールを人手で作成し、それらを用いて構造化文書の生成を行う方法がある。こうした方法には、次のような問題点がある。

1. 論理構造認識を行うための論理構造及びルールは対象となる文書の分野に依存するため、分野の異なる文書を扱う際には、その分野に対応したルールを新たに人手で作成する必要がある。

2. ある分野における複数種類の文書に対して共通性の高い単一のルールを用いるため、個別論理構造に固有の構成要素を直接認識できない。

筆者らは、文書の論理構造化を行う場合には、対象とする文書を活用するために文書の論理構造を規定する論理構造定義が設定されている場合が多いことに着目し、論理構造とそれを認識するためのルールを、個々の文書に設定された論理構造定義から変換して半自動的に作成し、そのルールに従った論理構造認識を行う構造化文書生成方法を提案する。この方法により、従来の手法における問題点が次のように解決できる。

1. 論理構造とそれを認識するためのルールを、個々の文書に設定された論理構造定義を変換して作成することで、論理構造の設計及びルール作成に要する労力を軽減できる。

2. 個々の論理構造定義を基に作成した構文解析ルールに従って構造化文書を生成するため、個別論理構造に即した構造化文書を直接生成できる。

3　SGML変換システム

図1のブロック図を用いてSGML変換システムの処理概要を説明する。

最初に、構文解析部の生成方法について述べる。まず、DTD修正部において、SGMLにおける論理構造定義であるDTD[2](図2(a))を非構造化文書の記述様式に沿うように修正し(図2(b))、その修正DTDと元のDTDとの差分情報をDTD差分データとして保持しておく。構文解析ルール生成部では、ルール変換規則を参照して修正DTDからyacc形式の構文解析ルール(図2(c))を作成する。そして、構文解析部自動生成手続き(yacc)を用いて、構文解析ルールから、構文解析ルールに記述された構文解析処理を実行する構文解析部(構文解析プログラム)を生成する。

次に、上記の方法によって生成した構文解析部を用いて、文字認識結果からSGML文書を生成する方法について述べる。キーワード抽出部では、キーワード抽出ルールを用いて、文字認識結果から論理構造を表現する文字列すなわちキーワードを抽出し、対象文書をキーワードとそれ以外の文字列とを要素とする集合として抽象化したキーワード/テキストモデルを生成する。構文解析部では、構文解析ルール生成部で作成した構文解析ルールに従ってキーワード/テキストモデルに対する論理構造認識を行い、論理構造を表すタグ情報をキーワード/テキストモデルに付与する。仮SGMLテキスト出力部で

SGML Conversion System for Regulation Document(2)
— From Charactor Recognition result to SGML Document —
†Yoshifumi SATO, Takuya OKAMOTO, Masatoshi HINO
‡Information System R&D Division, Hitachi Ltd.

[1]Standard Generalized Markup Language: 文書記述言語

[2]Document Type Definition: 文書型定義

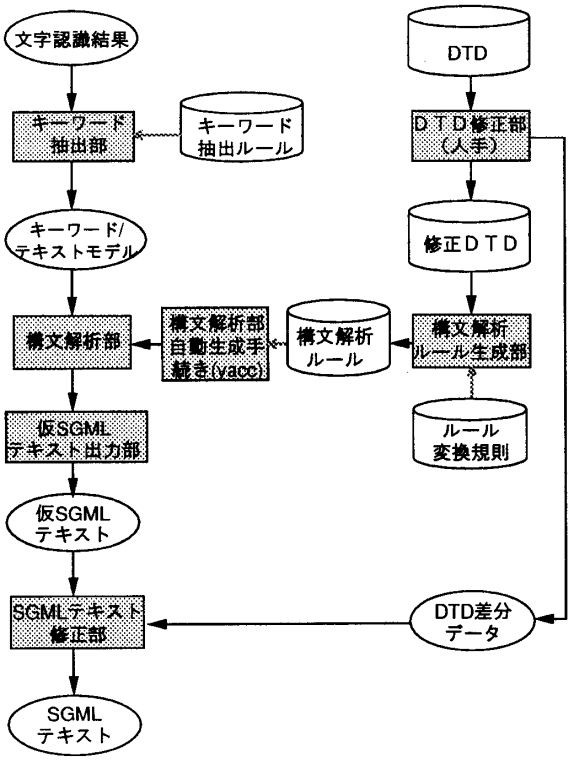

図 1: SGML 変換システムの処理概要

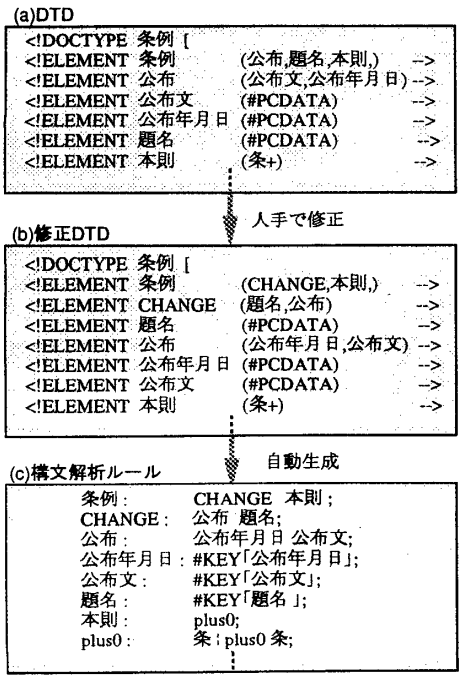

図 2: 構文解析ルール生成例

は、タグ情報付きのキーワード / テキストモデルを基に仮SGML テキストを生成する。これは、修正 DTD に沿った形で生成された文書インスタンスであるため、SGML テキスト修正部において、DTD 差分データを参照して仮 SGML テキストを修正することにより、元の DTD に沿った SGML テキストに変換する。

4 システム評価

4.1 評価実験

法規文書 114 件について、SGML 変換システムの評価実験を行った。評価結果を表 1に示す。85 %の法規文書に対して SGML テキストの生成に成功した。

表 1: 評価実験結果

総法規文書数	114 件
SGML 変換に成功した文書数	97 件
SGML 変換に失敗した文書数	17 件
SGML 変換成功率	85 %

4.2 課題の検討

SGML 変換に失敗した最大の原因は、文字誤認識によるキーワード抽出の失敗であった (17 件中 8 件)。これに対応するため、キーワード / テキストモデルの構文解析に失敗した時点で、その失敗した状況において本当はどのようなキーワードが存在すべきであったのかを調べ、構文解析に失敗した位置近辺の文字認識結果に対して、文字誤りを許容したキーワード抽出を再試行する手法を検討中である。

5 結論

本稿では、法規文書入力システムのサブシステムとして開発した、SGML 変換システムについて報告した。本システムは、対象とする文書に設定された論理構造定義 (DTD) から半自動的に生成した構文解析ルールに従う論理構造認識を行い、文書全体に論理構造を割り当てることにより、SGML 形式の文書を生成する。評価実験を行った結果、実験対象の 85 %の法規文書に対して論理構造認識に成功した。

参考文献

[1] 土居美和子ら. 文書構造抽出技法の開発. 信学論 D-II, Vol. J76-D-II, No. 9, pp. 2042–2052, (1993).

[2] 山田満. 文書画像の ODA 論理構造化文書への変換方式. 信学論 D-II, Vol. J76-D-II, No. 11, pp. 2274–2284, (1993.11).