プレインテキストの XML 形式への自動変換手法

丹野　志織里

XML(Extensible Markup Language)はデータ記述において汎用性に優れたマークアップ言語として普及しており，現在では多くの関連技術が存在する．例えば Web ページを記述する XHTML や，電子書籍の EPUB，オフィスソフト等のフォーマットに用いられており，XML の活躍する分野は多様である．

XML の長所の一つとして，ユーザーが自由に独自のタグを定義して利用できるということが挙げられるが，一方で，多様なタグを適切に付与していく作業には手間がかかる．そのため，タグ付けをサポートする機能が備わったエディタが複数存在する．既存の XML エディタの多くは半自動でタグ付けを行う機能を備えているが，全自動でプレインテキストにタグ付けを行えるという機能は，著者の知る限り存在しない．

そこで本研究では，プレインテキストに自動でタグ付けし，XML 形式に変換する手法を提案する．関連システムとして，プレインテキストから HTML 形式に自動変換する Markdown がある．Markdown は，プレインテキストに特定の記号を付与することや，決められた回数の改行を行うことなどにより，HTML への変換が行われる．それに対して，本手法は XML を対象としており，プレインテキスト作成時における制約は各行の改行とインデント数の指定のみである．

改行やインデントによりテキストの区切りを特定できるが，どのテキストにどのタグを付与するかは一般には一意に定まらない．このため，本研究では，プレインテキストに一意にタグ付けできるようなスキーマのクラスを定め，プレインテキストの XML 形式への自動変換を実現する．具体的には，スキーマの内容モデルの正規表現を 1-非曖昧性なものに限定し，型に優先度を付与することで一意なタグ付けを可能とする．このクラスに属するスキーマの各内容モデルに対してオートマトンを作成し，プレインテキストの各行をオートマトンの状態遷移に沿ってタグ付けを行う．ここで，内容モデルの正規表現は 1-非曖昧性なものであり，型に優先度が付与されていることから，遷移すべき状態が一意に定まり，一意なタグ付けが保証される．

評価実験では，２種類の XML Schema に準拠するプレインテキストを XML 形式に自動変換した．XML Schema は，授業メモと連絡先メモを想定したものを著者があらかじめ用意した．プレインテキストは，各 XML Schema に準拠するものを５名の協力者に作成してもらい，それぞれ XML 形式への自動変換を行った．その結果，適切に XML 形式への変換が行われることが分かった．このことは，用途に応じて適切なスキーマを定義すれば，プレインテキストから XML 形式への自動変換が可能になることを示している．

（指導教員　鈴木伸崇）