1Q-3

情報抽出を利用したNewsML編集支援システムの提案

浅川 浩紀 † 大園 忠親 ‡ 新谷 虎松 ‡
名古屋工業大学知能情報システム学科 † 名古屋工業大学大学院工学研究科情報工学専攻 ‡
e-mail: {asakawa, ozono, tora}@ics.nitech.ac.jp

1 はじめに

本稿では，情報抽出技術を利用した，NewsML 作成編集システムを提案する．

NewsML[1] はインターネットを通してニュースを配信する際に，ニュースの発信者，作成日および，ニュースの内容などのメタ情報を付加して，報道機関同士で情報を扱いやすくしたものである．主に報道機関同士のやりとりおよび，報道機関内でのニュース情報管理に使用されている．現在，数カ所で一般ユーザ向けの NewsML の公開も行われており，今後 NewsML 形式でのニュース公開が増えていくと思われる．NewsML には TopicSet 等非常に多くのメタデータを記述することが可能である．記事の種類，作成者，更新時刻，重要度，関連記事，提供者，記事の内容，場所等，非常に多くのメタ情報の記述が可能である一方で，これらメタ情報の付加は人手で行う必要がある．メタデータを有効に扱うには，メタデータ付加を補助する仕組みおよび，メタデータを自動的に付加する仕組みが不可欠である．

本研究では，NewsML 作成時のメタ情報付加の負担を軽減し，効果的な NewsML 作成を目的とする．本稿では，情報抽出による，ニュース編集時のインタラクティブなメタ情報付加補助システムおよび，既存のニュース記事を NewsML に自動変換システムを提案する．

本稿の構成を以下に示す．第 2 章では NewsML 編集支援システムの概要を述べる．第 3 章でメタ情報の自動抽出手法について述べ，第 4 章でニュース記事の NewsML への自動変換について述べる．最後に第 5 章で本稿をまとめる．

2 NewsML 編集支援システムの概要

本稿で提案する NewsML 編集支援システムの概要を述べる．本システムは，PHP，MySQL および，CaboCha[2] によって実装されている Web アプリケーションである．Web アプリケーションであるため，インターネットが利用できる環境であれば，どこからでもニュース記事の投稿および，編集が可能であり，リアルタイムにニュースを配信することが可能である．

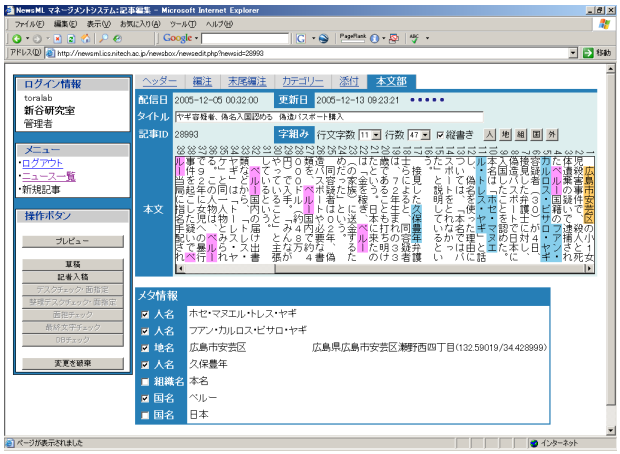

図 1: NewsML 編集支援システムの概要

ユーザ認証により，ユーザの権限をきめ細かく設定できる．また，ニュースデータは世代管理されており，差分をいつでも確認できる．ニュース記事には，画像だけでなく様々なコンテンツを付加することが可能である．校正者による記事修正，編集メモ機能，ニュースのリンク機能および，資料添付機能などにより，ニュース記事編集を支援する．ニュースは縦書きでの入力が可能で，ブラウザでプレビュー表示を行うこが出来る．編集中のニュースは即座にインデクシングされ，高速な全文検索が可能である．また，過去に編集したニュース記事を元に，関連ニュース記事の推薦を行う．

本システムでは，ニュース記者がニュース記事入力中に，ニュース記事から人名および，位置情報などのメタ情報を抽出する．ニュース入力者にとって，なるべく負担にならないメタ情報付加を実現するために，Ajax[3] 技術を利用している．ニュース記事入力中に，インタラクティブに自然言語処理により情報抽出を行い，抽出したメタ情報候補を提示する．ニュース記事作成者は，提示されたメタ情報をニュースに付加するかの判断および，メタ情報の修正を行う．

NewsML authoring system using information extraction
Hiroki ASAKAWA, Tadachika OZONO, and Toramatsu SHINTANI
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555 JAPAN

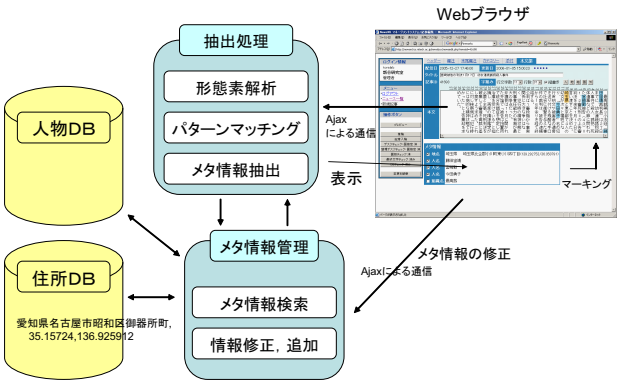

図 2: メタ情報抽出の流れ

3 メタ情報の自動抽出手法

図 2 に，メタ情報抽出の流れを示す．ニュース記事入力中に，Ajax により非同期に通信を行う．メタ情報抽出は，まず，ニュース記事に対して CaboCha により形態素解析および，固有表現抽出を行う．次に形態素解析結果に対して，パターンマッチングを行い，メタ情報付加の対象となる単語列を抽出する．例えば固有名詞-地域が連続していれば，住所として単語列を抽出する．抽出した単語列の種類に応じて，メタ情報ＤＢを検索し，メタ情報付加の為の情報を生成する．住所の場合，住所ＤＢ (住所と座標の対応表，街区レベル位置参照情報[1]を利用) を検索し，住所から座標データを得る．住所と座標データの組を位置メタ情報として，ブラウザに返す．人名の場合，人名を一意に特定する URN を人物ＤＢから検索する．人物名と URN の組をメタ情報としてブラウザに返す．抽出したメタ情報は，記事入力フォームの下に表示される．メタ情報の確認が行いやすいように，住所には地図へのリンク，人名には，同一人物が参照されている記事へのリンクが付けられる．メタ情報の抽出対象となった単語をマーキングする．正しく抽出されなかった場合には，メタ情報の修正が可能である．ニュース記事中の単語をマウスで選択し，抽出対象が人名であれば,「人」ボタンをクリックし，抽出を指示することで，再抽出が行われる．メタ情報管理機構により修正結果が蓄えられ，辞書を強化することで抽出精度が向上する．

4 ニュース記事の NewsML への自動変換

既存のニュース記事資源を NewsML 文章に変換するシステムについて説明する．まず，ニュース記事からパターンマッチングにより，ニュースの見出し部分，本文部分などを抽出する．抽出したニュース記事本文部分を CaboCha により形態素解析および，固有表現抽出を行う．CaboCha による解析結果に対して，パターンマッチングにより，メタ情報抽出元となる単語列を抽出する．抽出した単語列の種類に応じて，メタ情報 DB を検索し，メタ情報付加の為の情報を生成する．そして，メタ情報を付加した NewsML 文章を作成し，格納する．

次に，住所抽出について説明する．CaboCha 解析結果に対して，パターンマッチングにより，住所部分を抽出する．抽出した住所を住所 DB から検索し，座標データを得る．この際次のような処理を行い，抽出した住所の正規化を行う.「同県」,「同市」のような表現が含まれる場合，その単語出現以前に登場した県名により，参照を同定する．番地名の正規化を行う．一つのニュース記事中で，一番詳細な住所を住所 DB から検索する．見つからない場合は，住所を町名の前で分割し，AND 検索を行う．これはニュース記事において群名が省略される為である．見つからなかった場合は，次に詳細な住所を住所 DB から検索する．以下，座標データが得られるまで繰り返す．住所が都道府県市町村名までで終了していた場合，都道府県庁もしくは，市町村役場の座標データを付加する．

5 おわりに

本稿では，自然言語処理技術を利用した情報抽出による，NewsML メタ情報自動抽出が可能な，NewsML 編集支援システムを提案した．本システムを利用することにより，NewsML など，セマンティック Web における最大の問題である，メタ情報付加が半自動的に行われるため，効果的な NewsML によるニュース記事作成が可能になる．また，既存のニュース記事を NewsML に自動で変換することにより，ユーザにとって負担にならずに NewsML への移行が可能になる．

参考文献

[1] 井上明，猪狩淳一，金田重郎：ニュース配信のための国際データフォーマット NewsML:その概要と現状について，情報処理学会論文誌，Vol.2002，No.056，2002

[2] 工藤拓， 松本裕治: チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌，Vol.2002，No.043，2002

[3] Jesse James Garrett: Ajax: A New Approach to Web Applications
http://www.adaptivepath.com/publications/essays/archives/000385.php

[1] http://nlftp.mlit.go.jp/isj/