2B-5

# 演繹学習に基づく言語グリッドのラッパー生成

田仲 正弘 [1]　　石田 亨 [1]
京都大学大学院 情報学研究科 社会情報学専攻 [1]

## 1. はじめに

Web には HTML で記述された辞書や対訳集などの言語資源が存在する．図 1 はそのような言語資源の例であり，ユーザのクエリに応じて特定分野の用語の日英の対応やカテゴリ・関連語を返す．しかしこのような Web 上の言語資源は通常コンテンツにアクセスするための標準的なインターフェースを持たず，複数の言語資源を連携させて利用することが難しい．


図 1: ライフサイエンス辞書
(http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/)

本研究では Web 上の言語資源を，ページ構造の解析に基づいて，与えられたスキーマに適合する RDF/OWL による記述に変換する手法を提案する．これにより，言語資源の Web ページのための，標準的なインターフェースを持つラッパーを生成できる．生成されたラッパープログラムは Web サービスとして言語グリッド [1] に登録することにより，他の言語資源との連携を行える．

しかし，従来の Web ページからのラッパー生成手法をこのような RDF/OWL 記述への変換に適用するには次のような問題がある．

- ページ構造の自動的な解釈が難しい
- 訓練例作成のコストが大きい

従来 Web ページのラッパー生成手法として，タグのパターンなどに基づく自動的な解析を行う手法 ([2] など) が提案されてきたが，タグ構造のセマンティクスを自動的に正しく解釈することは難しい．また，訓練例からの帰納学習を用いる手法（[3] など）では訓練例の作成コストが大きく，言語資源 Web ページの提供者や簡単にその言語資源を他の言語資源と連携させたいと考えるユーザには困難である．

これらの点を解決するため，本研究では，演繹学習を用いてラッパーを生成する手法を提案する．初めに少数の訓練例を与え，領域知識に従って訓練例に対する説明を生成する．次に，生成された訓練例の説明を用いて未知のデータを解釈する．

## 2. ルールの生成

本研究では，ページ構造に関する領域知識として，与えられたスキーマに基づくルールを生成する．ルールは以下のような形式を持つ．

INSTANCE →
　$property\_value_1, \cdots, property\_value_n$

大文字の記号は非終端記号を示し，小文字の記号は終端記号を表す．$property\_value_1, \cdots, property\_value_n$ は，$P_1, \cdots, P_n$ のプロパティ値を表す終端記号である．このルールは，インスタンスからそのプロパティ値への展開を表す．

実際に用いるルールは以上の形式に従い，クラスごとに RDF スキーマから自動的に生成される．例として図 1 の Web ページを処理することを想定し，図 2 のような RDF スキーマが与えられるものとする．

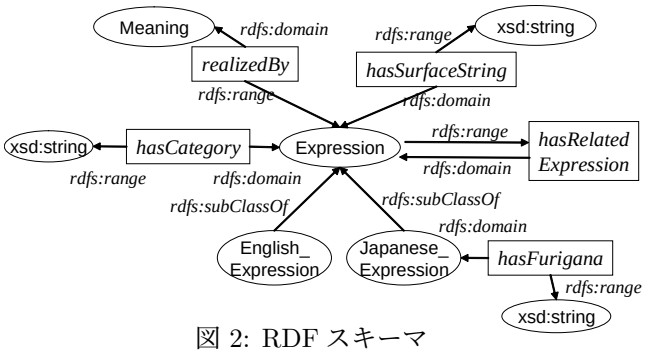


図 2: RDF スキーマ

このとき，クラスのインスタンスに関するプロパティの制約から，“Expression” クラスに関する以下のようなルールが生成される．

EXP_INSTANCE →
　surface_string | category | EXP_INSTANCE
　| surface_string category
　...
　| surface_string category EXP_INSTANCE

Wrapper Generation based on Deductive Learning for Language Grid
[*1]Masahiro Tanaka, Toru Ishida. Department of Social Informatics, Kyoto University

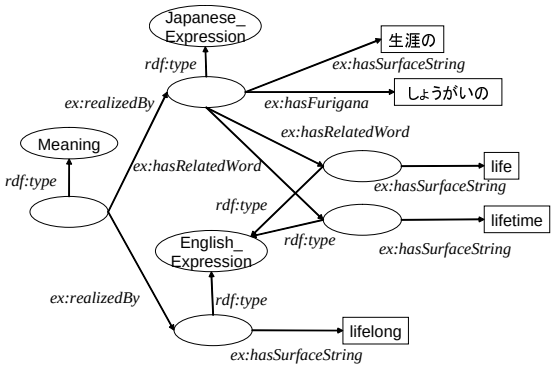


図 3: 訓練例

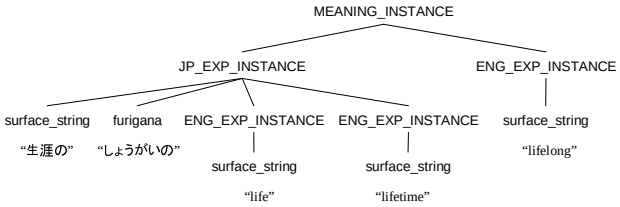


図 4: 図 1 のページ構造を表す説明木

## 3. ページ構造の解釈

Web ページの構造を解釈するため，ルール・目標概念・訓練例に基づいてページ構造の説明を得る．ここで目標概念とは，獲得されるべき概念のことを意味し，前章で述べたルールで用いられる記号で表現する．また訓練例は，図 3 のように，目標概念の記号が表すリソースと，そのリソースと関連づけられたリソースを RDF を用いて表現したものとする．それらから得られる説明は，ページ構造をルールに従って階層を木構造で表現したものであり，葉ノードに訓練例の RDF に含まれるテキストに相当する記号を持ち，根ノードに目標概念の記号を持つ．

説明の生成は，探索によって目標概念を根とする木構造を求めることで行う．このとき，各ノードに相当する Web ページ上の領域を，領域の包含関係によって求める．すなわち親ノードに相当する領域は，子ノードに相当する領域を含む領域とする．さらに，ノードに相当する領域を求めるデリミタを，[3] で示された手法に基づいて求める．図 1 のページに対して得られる木構造を図 4 に示す．これを説明木と呼ぶ．

次に説明木と同じ構造を，デリミタに基づいてノードを特定することによって，Web ページ中の別の場所で発見することを試みる．説明木に含まれるノードが見つからない場合には，類似のルールで同じように目標概念を導けないかを調べる．また，説明木に含まれない情報がある場合，その情報が何であるかをユーザが指示してやることで，木構造を修正する．このようにして，構造が未知の部分について得られる木構造を解釈木と呼ぶ．

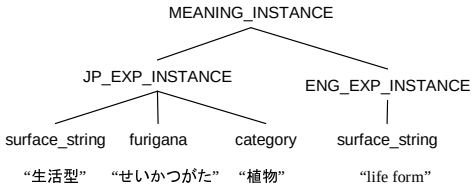


図 5: 修正された解釈木

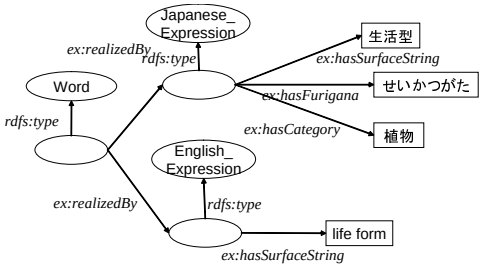


図 6: 図 1 から得られる RDF グラフ

図 1 中の訓練例として記述したのとは別の部分に対して得られる解釈木を図 5 に示す．この解釈木の構造から，訓練例とは別の語についての関係を表す RDF グラフが得られる（図 6）．

## 4. おわりに

本研究では，演繹学習を用いて HTML で記述される Web ページを RDF/OWL による記述に変換する手法を提案した．この手法は少数の訓練例で実行可能であり，また人間によるページ構造の解釈を利用できる．同じ形式のデータの繰り返しや獲得される概念間の関係と Web ページ上の領域の包含関係を利用するため，ある程度階層的なまとまりを含むページに適用できる．

これらの特徴から，本研究の手法は Web 上で提供される言語資源を標準的なインターフェースを持つ Web サービスとしてラッピングし，他の Web サービスと連携させる目的に有用であると考えられる．

## 参考文献

[1] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration, *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)*, pp. 96–100 (2006).

[2] Chia-Hui, C. and Shao-Chen, L.: IEPAD: Information Extraction based on Pattern Discovery, *Proceedings of the Tenth International World Wide Web Conference*, pp. 681–688 (2001).

[3] N. Kushmerick, D. S. W. and Doorenbos, R. B.: Wrapper Induction for Information Extraction, *Proc. of the 15th International Joint Conference on Artificial Intelligence*, pp. 729–737 (1997).