LMFに準拠したWordNet型意味辞書アクセスのためのWebサービス

Bora Savas，林 良彦（大阪大学大学院言語文化研究科）
{bsavas@gs.lang.osaka-u.ac.jp, hayashi@lang.osaka-u.ac.jp}

1. はじめに

本論文ではWordNet型の意味辞書にアクセスするためのWebサービスAPIについて提案する．REST (Richardson et al., 2007) の考え方に基づいたWebサービスを実現するためにWordNet型意味辞書へのアクセスパターンを検討し，それに基づくアクセスURI体系を設計した．意味辞書は，日本語版のWordNet 0.9 (以下WN-ja) (Bond et al., 2008)，Princeton WordNet 3.0 (以下PWN) (Fellbaum, 1998)および，EDR電子化辞書(以下EDR)(EDR, 2007)に対応している．

アクセスサービスは与えられたクエリに合致する部分的な辞書を抽出する処理であると考え，アクセス結果は辞書モデルに関する国際標準であるLMF (Lexical Markup Framework, Francopoulo et al., 2008; ISO 24613, 2008)に準拠したXML形式とする．

ただし，EDR電子化辞書のような複数言語にまたがる意味辞書を表現するには，そもそも単言語辞書を対象としているLMFには問題があることを示し，その拡張・修正を提案する．

2. WordNet型の意味辞書

2.1 WordNet型の意味辞書の定義

WordNet型意味辞書とは，情報構造として同義語集合であるsynsetと，上位(hypernym)/下位(hyponym)，全体(meronym)/部分(holonym)といった語彙意味論に基づく概念関係が記述されたPWNと同様の構造を持つ言語資源のことを示す．synset (synonym set)は，1つ以上の単語またはコロケーションからなる同義語の集合である．一般に単語は複数の語義を持つため，synsetの各要素は{word-form，part-of-speech，sense-number}という3つ組で規定される．

このような情報構造を持つ言語資源（wordnetとも呼ばれる）は様々な言語において開発され，Global WordNet Grid[1]においては，これらの統合が期待されている．

2.2 WordNet型意味辞書モデリングの枠組み

あらゆる言語の様々なタイプの辞書の情報構造をモデル化するための枠組み(メタモデル)として，LMF (Lexical Markup Framework) という国際標準が制定されている(ISO24613:2008)．EU KYOTOプロジェクト[2]により開発されたWordnet-LMF (Soria et al., 2009)は，特に各言語におけるwordnetにおいて定義された意味辞書の相互運用を実現するために設計されている．英語以外の言語でエンコードされた大規模な意味辞書の実例としては，日本語WordNet WN-ja (Bond et al., 2008)が公開された．

Wordnet-LMFでは二つ以上の異言語の辞書のエントリを対応付けるために，多言語拡張パッケージ(LMF Multilingual Notations)を用いている．多言語拡張パッケージにおけるSense Axisは，異なる辞書におけるsynset間の意味的な対応関係を表すものである．WN-jaにおける各synsetは一対一の等価的対応関係で結ばれるPWNのsynsetを持っている．

この手法は，それぞれ独立して開発された異言語の辞書を間接的に対応付ける．このため，多言語への展開において有効である．

2.3 EDRをWordNet型意味辞書としてモデル化

EDR電子化辞書は，日本語と英語を対象とする単言語辞書(日本語，英語)，対訳辞書（日英，英日），概念辞書などの言語資源の集合体である．

EDRの論理的構成を**図1**に示す．EDR電子化辞書の各辞書におけるすべてのエントリは，概念識別子(CID: *concept identifier*)によって関係付けられている．概念識別子は，日本語，英語にまたがる複数の同義語の辞書エントリにより参照される．このため，概念識別子をキーとし

[1] http://www.globalwordnet.org

[2] http://www.kyoto-project.eu

てこれと関係付けられた単語の集合を求めることにより，疑似的に日本語，英語にまたがるsynsetを構成することができる．さらに，概念体系辞書においては，概念間の上位・下位関係が構造化されている．以上より，EDR電子化辞書は，形式的にはPWNと同様の情報構造を持ち，WordNet型の意味辞書としてのモデル化が実現可能である．したがって，本Webサービスにおいても，PWN/WN-jaへのアクセスと同じ枠組みで扱うことができる．

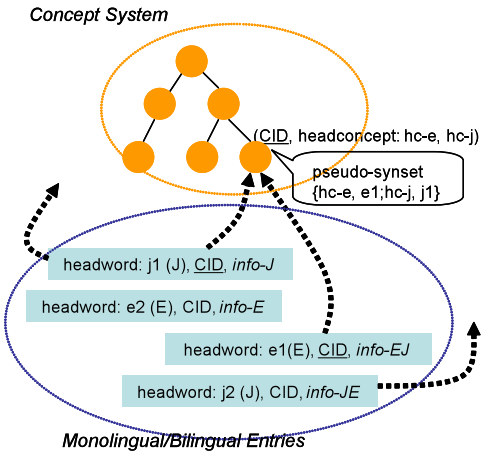

図1: EDRの論理的構成

3. RESTful Webサービス

Webサービスのメッセージ交換においては，主にWSDL/SOAPが利用されているが，本WebサービスではREST (Richardson, 2007) を利用する．REST形式ではSOAP[3]におけるヘッダーやエンベロープが不要であり，一般に扱い易い．また，サービス対象をリソースと考える点も辞書アクセスサービスに適している．

本WebサービスをRESTの考え方に基づいて実現するために，WordNet型意味辞書へのアクセスパターンを検討し，それに基づくアクセスURI体系を設計した．アクセスサービスは，与えられたクエリに合致する部分的な辞書を抽出する処理であると考える．RESTではWSDLを用いないので，結果データの解釈をクライアント側で独自に行う必要があるが，結果データを国際標準であるLMFに準拠させることにより，この負担を軽減できる．

3.1 APIおよび，その実例

辞書<Lexicon>はWordnet-LMFの仕様書に定義されているように，<LexicalEntry>と<Synset>の構成要素からなる．RESTの考え方に従えば，<LexicalEntry>は単語語形(以下の図2：Word)に対応するWebリソースであり，<Synset>はsynsetに対応するWebリソースであるため，**図2**で示すようにURI構造と対応するリソースをマッピングする．

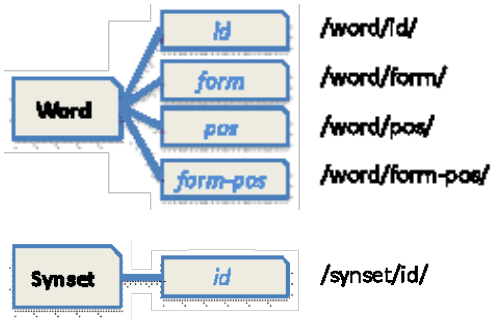

図2: URI構造および，リソースのマッピング

図2に示すように，指定された単語語形は対応する<LexicalEntry>要素の属性のどれかを用いてアクセスすることができる．ただし，'form-pos'属性は語形と品詞の属性の組合せである．

図3に，"銀行" (financial institute)に対応するWN-jaのエントリをLMFに準拠したXML形式で出力した例を示す．Webサービスは与えられたクエリに合致する部分的な辞書を抽出すると考えるため，XMLデータの最上階層の要素は<LexicalResource>であり，その直下の子ノードは<GlobalInformation>および<Lexicon>となる．

これに対し，Synsetリソースのプライマリなアクセスキーはidであるが，idは覚えるのが現実的ではない．そこで，ユーザに優しい形式でのアクセスを可能にするため，1）単語フォーム，2）単語フォームと形態素(POS)の組合せ，3）{単語フォーム，形態素，語義ID}の組合せによるアクセスを導入する．例えば，{"bank", n, 2} (語義：the financial institution)から対応するsynsetを検索する場合は，/word/form-pos/ のURIパターンに検索パラメータの(?synset_by_index)を指定する．

検索パラメータは，リソースの構造自体に直接関係していない制約を指定するのに便利な手法として導入する．また，単語フォームの語義説明文(以下gloss)に対して全文検索を行うための検索パラメータも導入している．例えば，/synset/?definition=Istanbulは，**図4**に示すようにglossに"Istanbul"を含むsynsetの集合を返す．また，前方一致や後方一致に対応した検索パラメータも指定できる．

3.2 実装

意味辞書言語資源の生のデータをRDBに格納し，各URIパターンに対応して生成されるSQL文により検索を実行

[3] http://www.w3.org/TR/soap12-part1/

する．バックエンドとしてはPythonで実装されたWebアピリケーションフレームワークDjango[4]を利用した．DjangoにはRDBとPythonの間の非互換なデータを変換するオブジェクト関係マッピング(ORM)と，柔軟にURI管理ができるURLマッピング機能が備えられている．また，出力データを生成するテンプレートシステムを利用することにより，Wordnet-LMFのXMLスキーマに準拠したXML文書を容易に生成できる．

4. Wordnet-LMF改訂の提案

LMFは，そもそも単言語の辞書資源をモデリングするための枠組みである．そのため，EDRのような複数言語を対象とする辞書のモデリングには問題がある．例えば，**図4** (id="edr_cph_1ee4d6-x") にあるように，EDRではglossが英語，日本語の双方で与えられる場合があるが，オリジナルのLMF，WordNet-LMFいずれも，glossは単一の言語で与えられることを前提としているため，これを素直に表現することができない．

LMFに変更を加えずに，このような情報構造を扱おうとすると，もともとは単一のレコードに収められていた情報を日本語側と英語側に分離し，それらを前述のSense Axisで結ぶといった工夫が必要となるが，これは本来的に複数言語の辞書を扱う方法としては不自然である．

そこで本稿では，<Synset>要素にlang属性で言語コードが指定した複数の<Definition>が含まれることを許すことを提案する．この改訂内容は元のLMFおよびWordnet-LMFに影響を及ぼさないものであり，生成されるすべてのXML文書はWordnet-LMFのスキーマで妥当性を検証することができる．

5. 関連研究

PWNにアクセスする最初のREST型のWebサービスは(Assem et al, 2006)によって開発され，RDF/OWLでコンテンツを表現するスキーマとそのURL設計が提案されている．本稿で提案するURL設計も彼らのURL設計の一部分を参考にしている．

{word-form}-{part-of-speech}-{sense-number}のようにハイフンで結合した3つ組によるsynsetの指定がその例である．

(Soria et al., 2006) は，PWNの構造に従って構築された各国語のwordnetがILI（Interlingual Language Index）により相互に接続される構造という連携形態において，各wordnetが提供すべきWeb APIの実装を提案した．また，Webサービスが返すsynsetのスキーマも提案しているが，このスキーマはLMFに準拠したものではない．

最近は，PWNおよびWN-jaにアクセスするためのAPIが言語グリッド[5]により提供されている．これらのWebサービスはREST型ではなく，WSDL /SOAPを利用している．アクセスの結果データはLMFに準拠した形式ではないが，WordNet型の意味辞書として(Hayashi et al., 2006)により提案されたEDRへのアクセスメソッドが提供されている．

6. まとめと今後の課題

本稿では，WordNet型の意味辞書にアクセスするためのREST型のWebサービスについて提案し，特にそのAPIについて述べた．意味辞書にアクセスするためのクエリを表現するURI構造について議論し，意味辞書の出力結果をWordnet-LMFスキーマに準拠した形式とすることを提案した．また，EDRのように本来が複数言語対応の辞書である言語資源に対応するため，<Synset>に複数の言語による<Definition>が含まれることを可能とするため，<Synset>要素にlang属性の言語コードを導入することを提案した．

以上の提案の有効性を実証するためには，今後，多言語のWordNet型意味辞書(wordnets)への適用性を確認していく必要がある．また，ユーザからの要求によって動的に複数の言語資源を組合せアクセスする場合の問題についても探求する．このような複合的なWebサービスを実現するためには，本サービスによる各wordnetのアクセスサービス，および，異なる辞書のエントリ間の対応付けサービスの実現が必要である．

謝辞

本研究は総務省戦略的情報通信研究開発推進制度(SCOPE) の援助を受けた．

本研究の一部は，ILC-CNR(イタリア)の Monica Monachini, Claudia Soria, Nicoletta Calzolariの各氏との共同研究による．

[4] http://www.djangoproject.com/

[5] http://langrid.nict.go.jp/langrid-developers-wiki-en/

参考文献

Richardson, L., and Ruby, S. 2007. *RESTful Web Services*. O'Reilly.

Francopoulo, G., Bel, N.,George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. 2008. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*, Vol.43, No.1, pp.57-70.

ISO 24613. 2008. *Lexical Markup Framework (LMF)*. ISO 24613:2008.

Bond, F., Isahara, H., Kanzaki, K., and Kiyotaka Uchimoto, K. 2008. Boot-strapping a WordNet using multiple existing WordNets. *LREC2008*.

EDR. 2007. The EDR Dictionary. http://www2.nict.go.jp/r/r312/EDR/index.html

Fellbaum, C. (Ed). 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Soria, C., Monachini, M., and Vossen, P. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. *IWIC 2009*, pp.139-146.

Hayashi, Y., and Ishida, T. 2006. A dictionary model for unifying machine readable dictionaries and computational concept lexicons. *LREC2006*, pp.1-6.

van Assem, M., Gangemi, A., and Schreiber, G. 2006. Conversion of WordNet to a standard RDF/OWL representation. *LREC2006*, pp.237-242.

Soria, C., Tesconi, M., Marchetti, A., Bertagna, F., Monachini, M., Huang, C.R., and Calzolari, N. Towards agent-based cross-lingual interoperability of distributed lexical resources. *COLING-ACL 2006 Workshop on Multilingual Language Resources and Interoperability*, pp.17-24.


```
−<LexicalResource>
  <GlobalInformation label="/wn-ja/word/form/銀行/"/>
 −<Lexicon languageCoding="ISO 639-3" label="Japanese Wordnet 3.0" language="jpn" owner="NICT" version="0.9">
  −<LexicalEntry id="w211859">
     <Lemma writtenForm="銀行" partOfSpeech="n"/>
     <Sense index="1" id="w211859_02787772-n" synset="jpn-09-02787772-n"/>
     <Sense index="2" id="w211859_13368318-n" synset="jpn-09-13368318-n"/>
     <Sense index="3" id="w211859_08420278-n" synset="jpn-09-08420278-n"/>
   </LexicalEntry>
  </Lexicon>
 </LexicalResource>
```


図3: XML出力: /wn-ja/word/form/銀行/


```
−<LexicalResource>
   <GlobalInformation label="/edr/synset/definition/Istanbul/"/>
 −<Lexicon languageCoding="ISO 639-3" label="EDR" language="en ja" owner="NICT" version="1.0">
  +<LexicalEntry id="edr_je_18910"></LexicalEntry>
  −<Synset id="edr_cph-1ee4d6-x" baseConcept="undef">
     <Definition lang="en" gloss="a city in Turkey, called Istanbul"/>
     <Definition lang="ja" gloss="イスタンブールという,トルコの都市"/>
    −<SynsetRelations>
       <SynsetRelation target="edr_cph-444a77-x" relType="upper-of"/>
       <SynsetRelation target="edr_cph-10a9d6-x" relType="goal"/>
       <SynsetRelation target="edr_cph-1e85e6-x" relType="goal"/>
     </SynsetRelations>
   </Synset>
  </Lexicon>
 </LexicalResource>
```


図4: XML出力: /edr/synset/definition/Istanbul