複合的な語彙資源アクセスサービスの実現基盤

林 良彦

大阪大学大学院言語文化研究科

1. はじめに

言語資源を Web サービス化し，これらの組み合わせによる多様な言語サービスを可能とする技術基盤の確立を目指すプロジェクトが進められている．本論文では，このような言語サービス基盤において，複数の語彙資源(辞書言語資源)を組み合わせ的にアクセスする複合的なアクセスサービスの構成要素について検討する．とくに，意味・概念に基づいて異種の辞書エントリを動的に対応付けるシナリオにおいて，これらの対応関係を二次的な言語資源として構築するための枠組みについて論じる．さらに，このような枠組みにおける辞書の情報構造の標準的なモデリングに関して，ISO 国際標準である LMF (Lexical Markup Framework) の適用と拡張について検討する．

2. 言語データ資源の Web サービス化

コーパスや辞書などの言語資源の本質は，それ自体では他に対して働きかけを行うことのない受動的・静的なデータである．このような言語資源を言語処理ツール・システムと区別する意味で言語データ資源と呼ぶ．どのような形態にせよ，言語データ資源を利用するためには，これにアクセスする処理機能が必要である．アクセス処理を Web 上で一定の標準化された形式で利用可能な形とすれば，言語データ資源を Web サービス化 (Web servicize) することが可能となる．ひとたび，言語データ資源が Web サービス化されれば，これらを必要に応じて動的に組み合わせることにより，複合的な言語データ資源 (Calzolari, 2008) を仮想的に実現することが可能となる．むろん，これを可能とするためには，サービスの実行基盤が必要であり，このための取り組みが欧州における CLARIN (Vradi et al., 2008) や言語グリッド (Ishida, 2006) において進められている．さらに，言語資源・サービスの相互運用性のための共有基盤として，言語サービスにかかわる多様な要素を形式的に規定するためのオントロジー体系の検討も進められている (Hayashi et al., 2008).

3. 複合的な辞書の意義

辞書学の分野においても，辞書データの電子化や情報技術の発達を背景として，複合的な辞書の可能性が議論されている．たとえば Hartman (2005) は，各種の可能な組み合わせ (hybrid genres) を提示している．

言語学習に関連しては，二言語化辞書 (bilingualized dictionary) の有効性が議論されている．たとえば，L2-L1-L1 辞書 は L2 の単語に対して対訳辞書から得られる L1 における訳語を単に提示するのではなく，各訳語の L1 における単言語辞書における豊富なエントリ情報を同時に提示するものであり，L1 を母語とする L2 学習者が L2 の理解タスクにおいて有用であるとされている (Hartman, 1994). 一方, L2 での産出タスクにおいては，L1-L2-L2 辞書が有用であるとされる (Laufer and Levitzky-Aviad, T, 2006). ここでも単に L2 における訳語を提示するのではなく，L2 の単言語辞書で提示されるような語義や用法に関する情報が同時に提示される．

4. ケーススタディとシナリオ: EDR+WordNet

上記のような複合的な辞書を Web 上で仮想的に提供するための技術基盤を検討するために，EDR 電子化辞書における概念体系辞書を L1(日本語)辞書,Princeton WordNet を L2(英語)辞書, EDR 電子化辞書の対訳辞書を L1-L2 対訳辞書とする L1-L2-L2 辞書をケーススタディとし，その実現シナリオを検討する．

4.1 EDR, WordNet の情報構造

WordNet の情報構造は語彙化概念に基づいている．WordNet における基本的な情報単位は synset と呼ばれる同義語集合である．単語はいくつかの語義を持ち，各語義はある概念を指示する．”car”, “automobile”のように異なる単語が共通の概念を指示する場合，これらの単語が synset を構成する．synset には自然言語テキストによる説明(gloss)が与えられる．ただし，gloss はあくまで語彙化概念の説明であり，その規定は他の synset との関係によって与えられる．ここで中心的な役割を果たすのは，上位/下位，全体/部分といった語彙意味論に基づく概念関係である．

EDR 電子化辞書は，日本語と英語を対象とする単言語辞書，対訳辞書，共起辞書などの言語資源の集合体である．EDR 電子化辞書における特徴は，これらの辞書のすべてのエントリが日本語・英語にまたがる概念識別子によって関係付けられていることである．概念識別子は，日本語，英語にまたがる(あるいは，言語に依存しない)概念ノードを表し，日本語，英語による見出し語と概念説明テキストが付与される[1]. たとえば，`0f74e9` という概念識別子はおおむね「自動車」の概念を表し，”自動車”や”car” といった日本語，英語の単語に関する各辞書のエントリはこの概念識別子と関係付けられている．このため，概念識別子をキーとし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，概念の上位関係が構造化され

[1] ただし，実際の EDR 概念体系辞書においては，日本語・英語の概念見出し，日本語・英語の概念説明がすべて揃っている概念識別子は少ない．

ている．以上より，EDR 電子化辞書は，形式的には Princeton WordNet と同様の情報構造を持ち，synset と gloss が日英両言語により与えられるものとして扱うことができる（Hayashi and Ishida, 2006).

4.2 サービスの実現シナリオ

複合的な語彙資源は，あらかじめ off-line の batch 的な処理により実現することもちろん可能である．実際，多くの語彙的オントロジーの対応付けに関する従来研究は，このような前提に基づいている．しかしながら，2 節で議論したように，言語データ資源の Web サービス化は動的で仮想的な言語資源の実現の可能性をひらく．そこで，ユーザからの要求によって(on-demand)，動的に(on-the-fly)，複数の語彙資源を組み合わせアクセスすることにより，仮想的な複合辞書へのアクセス機能を提供するアクセスサービスの実現を考える．すなわち，異なる辞書のエントリ間の対応付けをユーザからの要求に即して機会主義的に(opportunistic) 行う一方，対応付けの結果を Web サービス基盤上に新たな言語資源として蓄積する．

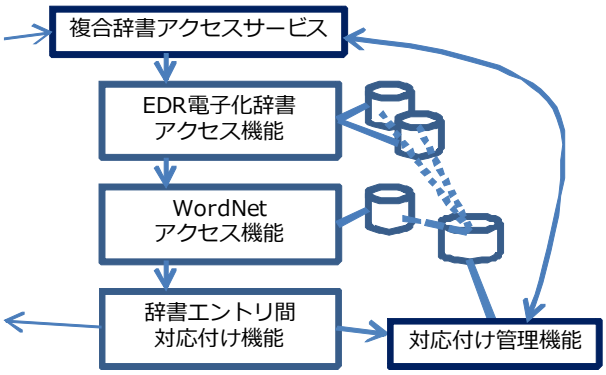

図 1: 複合辞書アクセスサービスの構成

このようなアクセスサービスの処理構成を**図 1** に示す．まず，すでに日本語のクエリ単語の各語義と WordNet が対応付けられているかを調べる．もしまだ対応付けが蓄積されていなければ，日本語のクエリ単語に対応する英訳語を EDR 日英対訳辞書から検索する．英訳語は一般に複数得られるが，各訳語には概念識別子が付与されている[2]．そこで，この概念識別子をキーとして得られる語義説明などの情報を手掛かりとして用い，それぞれに対応する WordNet の synset を求め，ユーザに提示する．さらに，対応付けの結果を蓄積する．言語処理的には，EDR 体系における概念を WordNet 体系の概念と対応付ける処理が必要となる．

4.3 概念による辞書エントリの対応付け

異なる語彙的オントロジーの対応付けに関しては，多くの先行研究がある．実際に EDR と WordNet の対応付けに関する研究も報告されている（Utiyama and Hasida, 1997)．とくに本研究における要求条件は，on-demand /on-the-fly /opportunistic な対応付けであり，ユーザに対して実時間程度で応答を返すことのできる負荷の軽い処理が求められる．そこで，英語における synset 単語集合の類似度，および，英語の概念説明テキストの類似度により対応付けを行う方法を適用する．より具体的には，これらにおける語彙のオーバーラップを評価する指標を統合的に用いる．なお，EDR 辞書において英語の概念説明が欠落している場合は，日本語の概念説明を機械翻訳により英語化した英語テキストを用いる．このような対応付けの詳細については，稿を改めて報告するが，

1. そもそも対訳辞書検索の段階で十分な英訳語が得られていないと，WordNet 側の検索が適切に行われないこと（再現率の問題)
2. 表層的なマッチング処理に基づくため，必ずしも適切な対応関係が高く評価されるとは限らないこと（精度，信頼度の問題)
3. さらに，得られる対応付けは概念的な等価関係とは限らないこと（対応関係の種別の問題)

といった点が問題となる．以上において，とくに 2, 3 の問題は，得られた対応関係をどのように記述すればよいかという情報構造のモデリングの問題と関連する．

5. 語彙資源モデリングの枠組み

5.1 LMF (Lexical Markup Framework)

LMF は，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)を規定する ISO 国際標準(ISO24613:2008) である．LMF の全体構造は，すべてのタイプの辞書に共通する Core package と各種の辞書を規定するための Extension packages から構成される．LMF の仕様は UML により規定されている．

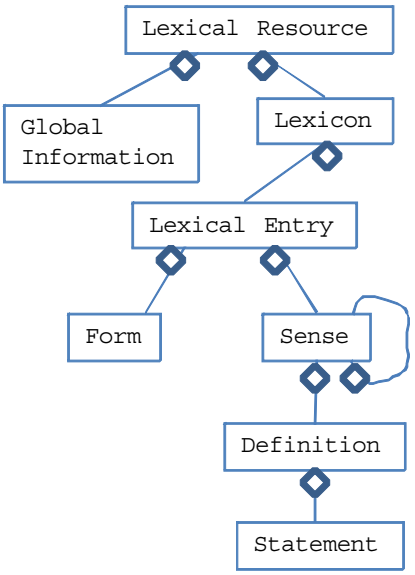

図 2: LMF Core package の構造

[2] より一般的な対訳辞書を用いる場合は，EDR のような語義説明が付与されているとは限らないため，そこで得られる情報の範囲での対応付けが必要となり，より困難な問題となる．

図 2 に Core Package の UML 図を示す．辞書エントリ（**Lexical Entry**）は，形式に関する情報（**Form**）と意味に関する情報（**Sense**）から構成される．

5.2 LMF における Sense Axis クラス

LMF の各 Extensions は，各種の辞書の既定に必要な情報項目を表現するため，Core package 中の必要なクラスを詳細化する．本稿の内容にとくに関係するものは，NLP Multilingual Notations（以下，多言語パッケージ）と呼ばれるパッケージである（Francopoulo et al., 2006）．多言語パッケージは，二つ以上の異言語の辞書のエントリを対応付けるため，NLP Syntax，NLP Semantics という二つの Extensions をさらに拡張するものである．多言語パッケージにおいては，異言語の辞書エントリの対応付けのために，**Sense Axis** と **Transfer Axis** という二つのクラスを導入している．これらはそれぞれ，機械翻訳方式におけるピボット方式とトランスファ方式に対応しており，**Transfer Axis** は，とくに二言語間の統語的な対応関係の表現に適している．一方，**Sense Axis** は意味的な対応関係を表すもので，とくに 3 つ以上の多言語間の等価的対応を表現するのに適している．

図 3 に **Sense Axis** に関連する UML 図の抜粋を示す．図に示すように，**Sense Axis** のインスタンスは，複数の **Sense/Synset**[3] のインスタンスと関係づけられる．これにより，多言語の異なる辞書の具体的なエントリ間の意味的な対応関係が表現できる．**Sense Axis** はさらに，**Sense Axis Relation** のインスタンスを集約し，**Interlingual External Ref** のインスタンスと関連する．前者は，辞書エントリ間の対応関係のさらなる関係（例：あるエントリ間の対応関係が別のエントリ間の対応関係より一般的である）を表現する．後者は，言語非依存な意味概念体系(いわゆる上位オントロジーや形式的オントロジー)とのリンキングを表現する．本稿の文脈において特に重要な点は，**Sense Axis** のインスタンスを集約したものも **Lexical Resource**（語彙資源）とされることである．これは，次節で議論する二次的な語彙資源の考え方を支持する．

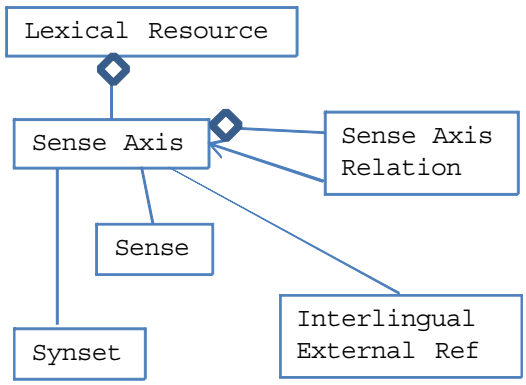

図 3: LMF における Sense Axis クラス

[3] **Synset** はとくに WordNet タイプの語彙資源を意識して NLP semantic extension で導入されたクラスである．

6. 二次的な言語資源

EDR 電子化辞書や Princeton WordNet は，独立した一次的な言語資源である．これに対し，**Sense Axis** インスタスを集約する語彙資源は，一次的な言語資源における情報間の関連を保持するという意味で，二次的な言語資源である．このような二次的な情報構造は，完全に一次的言語資源の外部にあるべきである．すなわち，一次的言語資源には，二次的言語資源への参照は含まれるべきではない．これにより，一次的言語資源は独立に保ったまま，その上に二次的な言語資源を重畳させることが可能となる．

6.1 LMF Sense Axis クラスの拡張

Sense Axis インスタンスを集約した二次的な言語資源を Web サービス化したものをここでは仮に Sense Axis サーバーと呼ぶ．異なる辞書エントリの間の対応関係を on-demand/on-the-fly/opportunistic に求め，その結果を保持するという本研究の文脈においては，Sense Axis サーバーは，関連付けられた特定の辞書エントリの組に関する情報も保持する必要がある．そこで，

- 対応関係：等価関係以外の語彙意味論的関係による対応付けがありえるため，その関係ラベルを保持する．
- 対応関係付与に関するメタ情報：いつ，どのようなプロセスにより，どの程度の信頼度をもって当該の関係が付与されたかを記録する．

といった情報を保持する **Sense Pair Relation** を導入し，Sense Axis サーバーにおいては **Sense Pair Relation** のインスタンスを対応する **Sense Axis** インスタンスと関係づけて集約することを提案する．**Sense Pair Relation** のインスタンスには，対応付けを保持している **Sense/Synset** インスタンスの ID と関連する **Sense Axis** インスタンスの ID を保持するとともに，対応付けにおけるメタ情報を保持する．**図 4** にこれらのインスタンスの関係の概要を示す．

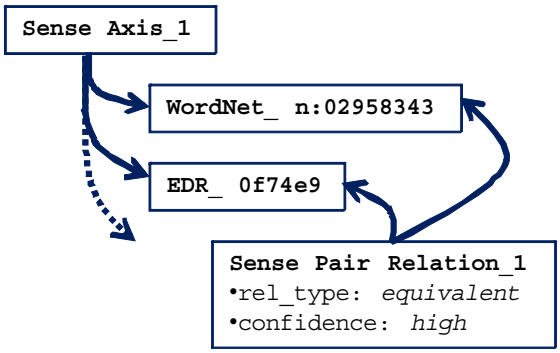

図 4: 各クラスのインスタンス間の関係

図 5 に LMF の **Sense Axis** クラスの拡張の概要を示す（図 3 からは **Sense Axis Relation** と **Interlingual External Ref** を省略）．ここでは，**Sense Axis Server** を **Lexical Resource** のサブクラスとして規定している．

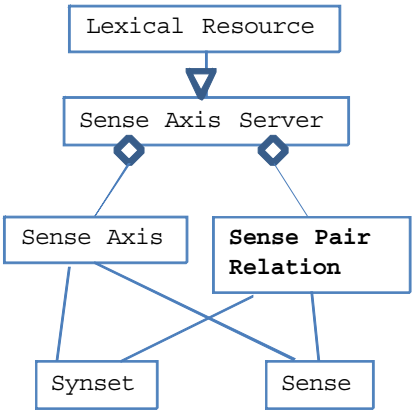

図 5: Sense Axis クラスの拡張

6.2 Sense Axis サーバーの Web API

Web サービスの形態としていわゆる REST（Pautasso, 2008）を想定する．すなわち，サービスの機能と対応した一定の URI 形式によりサービスにアクセスし，結果データを XML 形式で受け取る．本研究のベースは LMF にあるので，結果データの形式は LMF の仕様に annex として提示されている XML 形式に準拠する．そこで焦点は，Sense Axis サーバーが有するべき機能と URI の設計となる．なお，WordNet や EDR などの語彙資源が Web サービス化されていることを前提とする．すでに，(Van Assem et al., 2006) は，REST 形式での WordNet サービスを提案している[4]．また，REST 形式ではないが，言語グリッドでも両資源が Web サービス化されている[5]．

Sense Axis サーバーが有すべき基本機能は次のとおりであり，これらの機能との対応が明確な URI を割り当てる．

A) エントリ間の対応付けの検索と実行: 単語とターゲットとする語彙資源を与え，すでに辞書エントリ間の対応が得られていれば，その **`Sense Axis`** インスタンスの ID を返却する．対応付けが得られていなければ，動的な対応付け処理を行い，その結果を返す．本 API は，図 1 に示したアクセスサービスのトップレベルに相当する．

B) エントリ間の対応付け情報のアクセス: **`Sense Axis`** インスタンスの ID から関係付けられた **`Sense/Synset`** インスタンスの ID のリストを返却する．ここで，これらのインスタンスは各語彙資源における具体的なエントリであり，特定の URI 形式によりアクセスできるものとする．また，その辞書エントリの内容は，LMF/XML で得られるものとする．さらに，**`Sense/Synset`** インスタンスの組に関する **`Sense Pair Relation`** のインスタンスの ID のリストを返却する．

C) エントリ間の対応付け情報の登録: A の API による動的な対応付けの結果を適切なメタ情報とともに **`Sense Pair Relation`** のインスタンスに登録する．

[4] W3C でホストされており，rdf データを返却する．例えば: http://www.w3.org/2006/03/wn/wn20/instances/wordsense-bank-noun-1.rdf

[5] http://langrid.org/playground/concept-dictionary.html で試すことができる

なお，多言語の word nets の Web サービス化を想定した先駆的な研究に(Soria et al., 2006)がある．ここでは，各ローカルな wordnet に対して，ILI (Inter-Lingual Index; LMF での Sense Axis に対応)の ID からそれに関連する synset の ID を返す API を提案している．しかし，これは「二次的な言語資源は一次的な言語資源の外部にあるべき」という本研究の主張に反する．

7. おわりに

本稿では，異言語・異体系の意味・概念辞書を動的に対応付け，この対応関係を Web サービス基盤上で二次的な言語資源として蓄積する枠組みを検討し，辞書モデリングに関する ISO 国際標準 LMF の適用と拡張について議論した．今後は，Web API における URI 形式の詳細と結果の XML 形式を定め，実際に Web サービスの実現を行う．

参考文献

Van Assem, M., et al. 2006. Conversion of WordNet to a standard RDF/OWL representation. *Proc. of LREC2006.*

Calzolari, N. 2008. Approaches towards a “Lexical Web”: the Role of Interoperability. *Proc. of ICGL2008*.

Francopoulo, G., et al. 2006. LMF for Multilingual, Specialized Lexicons. *Proc. of LREC2006 Workshop on Acquiring and Representing Multilingual, Specialized Lexicons*.

Hayashi, Y., and Ishida, T. 2006. A Dictionary Model for Unifying Machine Readable Dictionaries and Computational Concept Lexicons. *Proc. of LREC2006*.

Hayashi, Y., et al. 2008. Ontologies for a Global Language Infrastructure. *Proc. of ICGL2008*.

Hartmann, R.R.K. 1994. Bilingualised versions of learners' dictionaries. In *Fremdsprachen Lehren und Lernen*. Gunter Narr Verlag, Tubingen. 23: 206-220.

Hartman, R.K.K. 2005. Pure or Hybrid? The Development of Mixed Dictionary Genres. *Linguistics and Literature*, Vol.3, No.2, pp.193-208.

Ishida, T. 2006. Language Grid: An Infrastructure for Intercultural Collaboration. *Proc. of SAINT2006*.

ISO 24613:2008. Language resource management - Lexical markup framework (LMF).

Laufer, B., and Levitzky-Aviad, T. 2006. Examining the Effectiveness of 'Bilingual Dictionary Plus' - A Dictionary for Production in a Foreign Language. *International Journal of Lexicography*, Vol.19, No.2, pp.135-155.

Pautasso, C. et al. 2008. RESTful Web Services vs. Big Web Services: Making the Right Architectural Decision. *Proc. of WWW2008*.

Soria, C., et al. 2006. Towards Agent-based Cross-lingual Interoperability of Distributed Lexical Resources. *Proc. of COLING-ACL 2006 Workshop on Multilingual Lexical Resources and Interoperability*.

Utiyama, M. and Hasida, K. 1997. Bottom-up Alignment of Ontologies. *Proc. of IJCAI-97 Workshop on Ontologies and Multilingual NLP.*

Vradi, T., et al. 2008. CLARIN: Common Language Resources and Technology Infrastructure. *Proc. of LREC2008*.