2B1-2

日本語リポジトリ「ことはぶ」の構築

Construction of Japanese Lexicon Repository “KotoHub”

丹 英之[*1] 大向 一輝[*2] 武田 英明[*2]
Hideyuki TAN Ikki OHMUKAI Hideaki TAKEDA

[*1]株式会社アルファシステムズ [*2]国立情報学研究所
Alpha Systems Inc. National Institute of Informatics

There are many types of dictionary and encyclopedia on the Web. Now we cannot find meaning of a word from these dictionaries at once, because their locations are distributed. In this paper, we produce a Japanese lexicon repository called “KotoHub”, which consists of millions of Linked Data written in RDF. Nodes of the repository are obtained from Web dictionaries and relations are from Japanese WordNet and NDLSH, pronouncing are from IPA dictionary.

1. はじめに

文書を作成する際，必要だと思われるものは辞書である．最近では，インターネット上の辞書検索サービスが充実してきたことで，直ちに語義を参照できるようになった．しかし，辞書サイトのサービスでは，そのサイトによって扱っている辞書コンテンツの編纂元が異なっており，語義を比較したい場合などには，検索行為を繰り返し，手間を要することとなる．そこで我々は，共通のインタフェース，標準化されたメタデータにより，複数のリソースを一元的に横断検索・参照できる日本語リポジトリ「ことはぶ」の構築を行った．本稿では,「ことはぶ」構築のために収集した言葉と，サービス層のインタフェースについて述べる．

2. Linked Data と「ことはぶ」

セマンティックウェブ技術の新しい潮流に，Linked Data[TBL 06] と呼ばれる概念が台頭してきた．ウェブの情報の価値は，そのコンテンツの内容と，他コンテンツとの繋がりの二つに依存しており，繋がりに相当するリンクが無ければ価値が減少する，という考え方がある．この繋がりを担保するため，データを RDF で表現し，外部から参照可能にしたウェブが Linked Data である．この，世界中にある，あらゆる RDF トリプルを互いに連携させようとする Linkd Data のために，4 つの条件が挙げられている．

1. 事柄の名前に URI を使うこと．
2. 名前の参照が HTTP URI でできること．
3. URI を参照したとき有用な情報が手に入ること．
4. 外部へのリンクを提供すること．

リポジトリの概念は古くから存在しているが,「ことはぶ」では Linked Data の概念を取り込み,

1. 全ての言葉の表記に URI を与え,
2. HTTP で参照でき,
3. 人間によるウェブブラウザからのアクセスのための HTML 表現と，エージェントによる計算機処理可能な RDF 表現のインタフェースを用意し,
4. 言葉の定義が記載されているコンテンツへのリンクを提供する,

言葉のハブとして機能するリポジトリ，“Dictionary of Dictionaries” を目指すことにした．

連絡先: 丹英之，株式会社アルファシステムズ，川崎市中原区上小田中 6-6-1，044-733-4111，tanh@alpha.co.jp

3. 言葉の収集

本章では，言葉の収集方法，そして言葉のソースとした，日本語 WordNet[Bond 09], 国立国会図書館件名表目標 [嶋田 07], IPAdic legacy, インターネット上の辞書サイトについて順に述べる．

3.1 言葉の収集方法

言葉である語を収集するには，大きく分けて二つの方法がある．一つは，情報工学的手法によって大量のコーパスを分析することで，語の文字列を発見・収集する方法である．もう一つは，辞書の編纂やオントロジの構築時に行われる，人間が一語ずつ確認し収集する方法である．

「ことはぶ」の要件定義は，言葉の定義元を保持し，それをサービスとして提供することである．そこで我々は，後者の手法によって集められた辞書の見出し語を語彙集合としてマージしていく手段を取った．

各辞書の見出し語を対象とした正規表現を用いることで，コンテンツから言葉を切り出した．得られた語は，括弧の削除，半角カタカナの全角化，全角英数文字の半角化，大文字の小文字化を行うことで正規化した．また，コンテンツに読みが記載されている場合は，読みも取得した．これらの処理でエラーが出たものは除外した．

3.2 日本語 WordNet

語は，概念を表す手段である．この概念とは，ある物事に対して共通事項を包括し，抽象・普遍化してとらえた意味内容のことであり，人間の思考活動の形態の一つに，概念間の関係操作がある．概念は語と連結しているので，概念同士の関係は，概念を表す手段の語の関係にも派生する．この関係を「ことはぶ」へ導入するには，語の関係性が記載された辞

書であるオントロジーが必要になる．そこで我々は，日本語 WordNet[Bond 09] に着目した．

日本語版 WordNet では，言葉である語の表記を表す Word エンティティ，語の表記と概念の連結にて，多義性，同義性を扱うために用意された中間テーブルの Sense エンティティ，そして概念を表す Synset エンティティによって構成されている．概念同士の関係を表す Sysnset 間の関係は，“See also”，“Synonyms”，“Hypernyms”，“Instances”，“Hyponym”，“Has Instances”，“Meronyms(-Member, -Substance, -Part)”，“Holonyms(-Member, -Substance, -Part)”，“Attributes”，“Similar to”，“Entailes”，“Causes”，“Domain(-Category, -Usage, -Region)”，“In Domain(-Category, -Usage, -Region)”，“Antonyms” の 23 種類が関係が定義されている．

「ことはぶ」では，この Synset 間の関係を用い，語間の意味リンクを構築した．日本語 WordNet には語の表記に英語も含まれているが，今後の語彙拡張を踏まえ，そのまま取り込むことにした．また，日本語 WordNet では，Synset エンティティから上位オントロジーである Suggested Upper Merged Ontology(SUMO)[Niles 01] への意味リンクが記載されている．そこで,「ことはぶ」からも SUMO の公開サイト[*1]へ外部リンクを提供することにした．日本語 WordNet のバージョンは 0.92 を用いた．

3.3 国立国会図書館件名表目標

意味リンク導入のため，オントロジーを利用することについては，前節で述べたとおりである．更にオントロジーを増強することで，語間に新たな意味リンクが生成され，辞書を跨いだ近傍単語の発見など，新しいユーザ体験が期待できる．そこで，階層構造をもつ国立国会図書館件名標目表 (National Diet Library Subject Headings, NDLSH)[嶋田 07] 2008 年度版 [国立国会図書館 08] を導入することにした．

NDLSH は，国立国会図書館の和図書・洋図書の目録において使用実績のある件名標目を収録した一覧表である．この件名標目を，言葉である語として扱った．件名標目は，主標目，細目，細目付き件名標目の 3 種類に別けられる．これら件名標目には，関係性の深い件名標目を参照できるよう,「も見よ参照」として「上位語」,「下位語」,「関連語」の 3 種類の関係を定義している．また，規約により主標目として使用できない件名標目については，注記によって細目に対し「を見よ参照」,「を見よ参照あり」の関係も定義している．

NDLSH では，実用性を伴う図書分類であることから，細目付き件名標目や限定語が括弧で付与された件名標目もある．そこで,「ことはぶ」では上記の基本関係 5 種類の他に，細目付き件名標目から,「細目付き件名標目の上位」,「細目付き件名標目の下位」,「細目付き件名標目の要素」,「を含む細目付き件名標目」を，括弧付き件名標目からは，括弧内の語を「限定語」，括弧付きの語を「限定語付与あり」とした関係の 6 種類を抽出し，これらも意味リンクとして加えた．

3.4 IPAdic legacy

言葉には，書き言葉と話し言葉の二つの側面がある．これらを,「ことはぶ」へ導入するには，書き言葉である語を表す文字列の記号体系と，発話に用いる音声符号とを相互変換するテーブルが必要になる．そこでこのテーブルを用意するため，語の読み方が記載されている IPAdic legacy(2.7.0) [*2] を用いることにした．

IPAdic legacy は形態素解析器 ChaSen の辞書として知られており，見出し語や読みの他に，形態素解析に必要な形態素生起コストと品詞情報が含まれている.「ことはぶ」では，見出し語と読みのエントリーを参照し，読みリンクを構築した．

3.5 インターネット上の辞書サイト

インターネットには，様々な辞書サイトが存在している．そこで，これらのうち扱い易いものを「ことはぶ」における語の定義先として採用することにした．扱い易さの基準は，次の 2 つである．

1. 語の定義が記載されいているコンテンツを，HTTP の GET メソッドで直接取得できること．
2. コンテンツを指示す URL のパスに，語の表記が含まれ，且つ，静的コンテンツである様に見えること．

特に，語の表記が含まれている URL は，ウェブブラウザの URL デコード表示機能により，一目で URL の指示すコンテンツの内容を把握できるので，非常に扱い易い．

各辞書サイトから得られる言葉の集合は，その辞書コンテンツの生成過程の違いから二つのタイプに分類できる．一つは，ユーザ参加型コミュニティによってマスコラボレーションされた知識体系から成る語彙集合，そしてもう一つは，専門家によって編纂された知識体系から成る語彙集合である.「ことはぶ」では，ユーザ参加型コミュニティにてボトムアップに生成・編纂された辞書として，Wikipedia(日本語) [*3]，はてなキーワード[*4]，ニコニコ大百科 (仮)[*5] を対象とした．また，専門家によってトップダウンに編纂された辞書として，小学館の「日本大百科全書 (ニッポニカ)」をベースに構築されている Yahoo!百科事典[*6]，小学館の「デジタル大辞泉」，日立システムアンドサービスの「百科事典マイペディア」，朝日新聞出版の「知恵蔵」をベースに構築されている用語解説サイトのコトバンク[*7]，三省堂の「大辞林」をベースに構築されている辞書サイト Weblio[*8]を対象とした．これにより「ことはぶ」では，異なる生成過程を辿った知識体系が融合されることになる．

各サイトをクロールして得られたコンテンツから見出し語を抽出し，その語の定義が記載されているコンテンツを，語の定義元とした．また，はてなキーワード，ニコニコ大百科 (仮)，Yahoo!百科事典からは，読みも取得した．なお，Weblio はコンテンツとして Wikipedia(日本) を包含しているため，索引ページのコンテンツを指す URL のパスを用い，Wikipedia(日本) からの登録語を除外した．各サイトのクロールは，2009 年 6 月から 2010 年 3 月までの期間に行った．

4. 収集した言葉

本章では，収集した言葉の語数と辞書の専門性，及び，辞書間における言葉の重複度について述べる．

4.1 収集した言葉の語数と辞書の専門性

3. で述べた辞書から，全部で約 295 万語の言葉を収集した．そのうちユニークな語は，約 225 万語であった．また，読みを取得できた語は，約 81 万語であった．語の収集元とした辞書名と，取得できた語数，及び，その辞書のみから取得できた

*1 http://www.ontologyportal.org/
*2 http://sourceforge.jp/projects/ipadic/releases/24435
*3 http://ja.wikipedia.org/
*4 http://d.hatena.ne.jp/keyword/
*5 http://dic.nicovideo.jp/
*6 http://100.yahoo.co.jp/
*7 http://kotobank.jp/
*8 http://www.weblio.jp/

表 1: 辞書名と取得語数，及び，その辞書のみから取得した語数とその割合

辞書名	取得語数	その辞書のみ	%
日本語 WordNet	231,865	152,304	65.9
NDLSH	51,093	20,155	39.4
IPAdic legacy	232,002	113,312	48.8
Wikipedia(日本語)	627,411	411,582	65.6
はてなキーワード	268,539	109,477	40.8
ニコニコ大百科 (仮)	48,939	26,484	54.1
Yahoo!百科事典	105,056	22,707	21.6
コトバンク	385,548	198,238	51.4
Weblio	953,941	810,494	85.0

語数とその割合を，表 1 に示す．語の比較では，3.1 で述べた正規化後の文字列を扱った．

最も言葉の定義が多い辞書は，Weblio であった．その辞書のみから取得できた語数をみると，Yahoo!百科事典は，約 11 万語の取得語数にも拘わらず，他の辞書に含まれない語は 22%と少なく，各分野のバランスがとれた語彙集合であると言える．一方，ニコニコ大百科 (仮) は，取得語数が約 5 万語で Yahoo!百科事典からの取得語数の半分以下であるが，他の辞書に含まれない語が 54%あり，専門性の高い語彙集合であると言える．これは，コミュニティを構成するユーザ層の偏りに由来するものであると考えられる．

4.2 辞書間における収集した言葉の重複度

各辞書の特性を見るため，辞書から取得できた語が，どの程度重なり合うかを調べた．各辞書について，その辞書から見た他の辞書との収録語の重なり具合を，表 2 に示す．

はてなキーワードと Yahoo!百科事典から見た，Wikipedia(日本語) に対する重複度は，それぞれ 44.3%, 45.5%であるが，はてなキーワードからの取得語数が約 27 語で Yahoo!百科事典からの取得語数が約 11 万語である．これより，はてなキーワードと Wikedia(日本語) は似た傾向をもつ語彙集合であると言える．また，ニコニコ大百科 (仮) では，Wikipedia(日本語) とはてなキーワードとの重複度が高い．一方，Yahoo!百科事典から見た，コトバンクと Wikipedia(日本語) に対する重複度は，それぞれ，56.9%, 45.5%であり，どちらの辞書とも重複度は高いが，コトバンクからの取得語数は約 39 万語であり，Wikipedia(日本語) からのそれは約 63 万である．これより，Yahoo!百科事典とコトバンクは極めて似た傾向をもつ語彙集合であると言える．これらの違いは，3.5 で述べた，各辞書の生成・編纂過程の違いに由来するものであると考えられる．

5. 「ことはぶ」のサービス層

本章では,「ことはぶ」のサービス層について，人間が直接利用する場合であるウェブブラウザからの利用，そして，エージェントによる計算機処理での利用を想定した，RDF 対応クライアントからのアクセスについて述べる．

5.1 ウェブブラウザからの利用

「ことはぶ」構築における目的の一つは，複数辞書の実用的な横断検索である．ユーザ・インタフェースは，検索クエリの入力欄と検索方法 (入力された文字列と読みの完全，部分，前方，後方一致) の選択，そして，検索実行のボタンだけであり，

図 1: 「ことはぶ」の検索結果画面

極めてシンプルな構成となっている．また，OpenSearch[*9] に準拠しており，ウェブブラウザにある検索ボックスから容易に検索が行える．

「ことはぶ」の検索結果画面を図 1 に示す．検索結果には，クエリとして入力された言葉に関係する，意味繋がり，読み繋がりによる「ことはぶ」自身の内部 URL リンク，及び，言葉の定義が記載されているコンテンツへの外部 URL リンクが提示される．ユーザは，それら URL リンクを辿ることで複数の定義を見比べることが可能となる．

「ことはぶ」では，各語の定義が記載されているコンテンツへのリンク集がコンテンツとなっており，その URL のパス部分は，“`/term/{URL エンコードされた語の文字列}`” となっている．URL エンコードされた文字列は，バイナリの ASCII 変換によってサイズが増加するため，Twitter[*10]に代表されるソーシャルメディアへの投稿には不向きである．そこで「ことはぶ」内コンテンツが一意に決まる短縮 URL を用意し，その URL による参照も可能にした．

5.2 RDF 対応クライアントからのアクセス

ウェブブラウザの Firefox では，Add-ons の The Tabulator Extension[*11]を導入することで，RDF ブラウザとして機能するようになる．これら RDF 対応クライアントでは，HTTP ヘッダの Accept リクエストフィールドの値を，“`application/rdf+xml`” もしくは “`text/rdf+n3`” としてリクエストを送信する．そこで，Accept リクエストフィールドの値を判別することでコンテントネゴシエーションを行い，RDF 対応クライアントからのリクエストの場合には，“`/term/{URL エンコードされた語}.rdf`” へリダイレクトを行う.「ことはぶ」は，この拡張子 “`.rdf`” の付いた URL へのリクエストに，RDF 表現のコンテンツを返す．この RDF 表現内では，その言葉の定義が記載されているコンテンツの URL リンクをプロパティ “`rdfs:seeAlso`” で示す．また，語の読みは，要素 “`linked:yomi`” を定義し，これを用いて「ことはぶ」内 URL リンクを示す．

我々は，RDF/OWL Representaion of WordNet[W3C 06] を WordNet3.0 用に変更 [小出 06] し，日本語 WordNet に適用した，日本語 WordNet の RDF/OWL 表現を参照できる別のサービス[*12] を稼働させており，このサービスが返す RDF

*9 http://www.opensearch.org/Documentation/Developer_how_to_guide
*10 http://twitter.com/
*11 http://dig.csail.mit.edu/2007/tab/
*12 http://wordnet.jp/

表 2: 辞書 A から見た辞書 B との収録語の重複度 (%)

	辞書 B								
辞書 A	日 WN	NDLSH	IPAd	Wiki 日	はてな	ニコ大	Ya!百	コトバ	Weblio
日本語 WordNet	100.0	5.4	15.7	11.2	7.8	1.5	6.3	20.9	13.3
NDLSH	24.3	100.0	22.7	27.9	22.2	4.1	28.7	38.7	26.6
IPAdic legacy	15.8	5.0	100.0	17.3	15.0	2.9	12.8	32.6	17.5
Wikipedia(日本語)	4.1	2.3	6.4	100.0	19.0	2.3	7.6	14.3	11.4
はてなキーワード	6.7	4.2	13.0	44.3	100.0	6.4	11.0	19.1	19.6
ニコニコ大百科 (仮)	7.1	4.3	13.6	30.1	35.3	100.0	8.8	13.6	17.2
Yahoo!百科事典	13.9	14.0	28.2	45.5	28.1	4.1	100.0	56.9	27.8
コトバンク	12.6	5.1	19.6	23.4	13.3	1.7	15.5	100.0	14.1
Weblio	3.2	1.4	4.3	7.5	5.5	0.8	3.1	5.7	100.0

内に「ことはぶ」への URL リンクを組み込んでいる．よって現状では,「ことはぶ」と日本語 WrodNet RDF/OWL 表現サービス間でのみ，相互参照可能な Linked Data として機能している．

6. まとめ

本稿では，インターネット上に散在している辞書・百科事典サービスの統合化を企図して構築した，日本語リポジトリ「ことはぶ」に関し，リポジトリ構築にあたり収集した語，そして「ことはぶ」のサービスについて議論した．

「ことはぶ」の存在により，異なる分野を跨いだ言葉の関係を把握できるため，言葉の探索において複合領域的なアプローチをとることが可能となった．

今後は，辞書の増強を行うと共に，様々なデータベースへの外部リンクを検討したい．辞書の増強については，他のデータベースに登録されているキーワードを取りこむことを検討している．これにより，データベース間を跨いだ語の参照が可能となる．日本語 WordNet には，英語表記の語が残っているので，これを手掛かりとして英語圏のリソースへの外部リンクが可能となる．これには，Wikipedia にある多言語リンクも利用できるので，日本語・英語圏以外のリソースへも到達可能であると考えている．また，NDLSH を導入したことから，書籍情報に関するリソースがあれば，それも外部リンクの対象と成り得る．更に，SKOS[W3C 04] を用いることで概念の階層構造を表現し，他のオントロジーへのリンクを用意することも検討したい.「ことはぶ」では，意味と読みにて内部 URL リンクを構成しているが，語の定義先である外部 URL リンクは，語の表記が一致することでの繋がりでしかない．これにより，検索したい言葉に多義性がある場合，意味を解釈することなく複数の定義先を提示してしまう．辞書の増強に合わせ，言葉の用いられていた文脈を考慮した，定義先の検索結果をランキング，もしくは分類する手法の導入が望まれる．これら課題を解決し，日本語圏における Linked Data のハブ，インフラストラクチャとなれるよう検討を進めたい．

構築した日本語リポジトリ「ことはぶ」は，

`http://wordnet.jp/kotohub`

にて公開している．

参考文献

[TBL 06] Tim Berners-Lee: Linked Data - Design Issue, `http://www.w3.org/DesignIssues/LinkedData.html` (2006)

[Bond 09] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki : Enhancing the Japanese WordNet, In The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009, Singapore. (2009)

[嶋田 07] 嶋田真智恵：国立国会図書館件名標目 (NDLSH) の改訂作業と今後について, 情報の科学と技術 57 巻 2 号, 73-78. (2007)

[国立国会図書館 08] 国立国会図書館, 国立国会図書館件名標目表 2008 年度版テキストデータダウンロード, `http://www.ndl.go.jp/jp/library/data/ndlsh_download.html` (2008)

[Niles 01] Niles, I., and Pease, A. : Towards a Standard Upper Ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems. (2001)

[W3C 06] W3C, RDF/OWL Representation of WordNet, W3C Working Draft 19 June 2006. (2006)

[小出 06] 小出誠二，森田武史，山口高平，ムリアディヘンドリー，武田英明：WordNet と EDR の OWL 表現, 第 13 回セマンティックウェブとオントロジー研究会 (2006)

[W3C 04] W3C, SKOS Simple Knowledge Organization System, `http://www.w3.org/2004/02/skos/` (2004)