様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　24 年　5 月　8 日現在

機関番号：14401
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21520401
研究課題名（和文）　語彙資源を用いた概念の語彙化の分析・記述に関する研究

研究課題名（英文）　A Research on Concept Lexicalization using Language Resources

研究代表者
林　良彦（HAYASHI YOSHIHIKO）
大阪大学・大学院言語文化研究科・教授
研究者番号：80379156

研究成果の概要（和文）：概念の語彙化の言語による相違は，多言語の語彙知識ベース・言語的オントロジーを構築する際に考慮すべき最大の課題である．そこで，ある言語の語彙概念を表す「単語」に対応する異言語の表現が語彙化されているかを判定する基準を明らかにし，さらに，語義タグ付きのコーパスと対訳資源を用いることにより，良好な精度で異言語の語彙概念を対応付けるための方法論を開発した．また，概念の語彙化の言語による共通性と差異性を表現するための情報構造モデルを提案した．

研究成果の概要（英文）：It is crucially important to understand potential differences in concept lexicalization among various languages. To tackle this linguistic issue and to develop a computational mechanism for constructing a useful multilingual lexical ontology based on existing lexical resources, the reported research developed a method to associate possibly corresponding lexicalized concepts across languages by using a sense-tagged corpus, and proposed a framework for representing a dynamically acquired cross-lingual/interlingual lexical-conceptual relations.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	900,000	270,000	1,170,000
2010 年度	1,000,000	300,000	1,300,000
2011 年度	1,200,000	360,000	1,560,000
総　計	3,100,000	930,000	4,030,000

研究分野：人文学
科研費の分科・細目：言語学・言語学
キーワード：辞書論，言語資源，語彙化概念，語彙ギャップ

１．研究開始当初の背景

概念の語彙化の言語による相違は，多言語の語彙知識ベースを構築する際に考慮すべき最大の課題である．そこで，本研究課題の申請時における背景・動機は，以下のように要約できる．

- ある言語(例えば日本語)の語彙資源における語彙概念と対応しうる,別の言語(例えば英語)の語彙資源における語彙概念を見出す計算的方法論を明らかにし，
- 対応付けが困難な語彙ギャップを収集・分析することにより，概念の語彙化に関する言語横断的な研究の課題を明らかにする．

これらに関する研究成果により，既存の各国語の語彙資源を組み合わせることよって，多言語の語彙知識ベース・言語的オントロジーを構築するための方法論・枠組みの構築に寄与し，語彙ギャップの言語横断的な研究のための言語資源を構築する．

２．研究の目的

以上より，本研究課題の申請時における当初の研究目的は，以下のように要約できる．

- ある言語の語彙概念を表す「単語」に対応する異言語の表現が語彙化されているかを判定する基準，および，手法を明らかにする．
- 与えられた２言語の語彙概念が意味的にどの程度対応しうるかを計量する計算的手法を開発する．
- 概念の語彙化の言語による共通性と差異性を表現するための情報構造モデルを明らかにする．

なお，本研究課題の実施においては，対象言語を日本語，英語に限定した．

３．研究の方法

(1) 語彙化を判定する基準・手法

特に語彙ギャップ（ある言語における語彙に直接対応する語彙が別の言語において存在しないという現象）に注目することにより，概念語彙化の言語による差異を分析・記述するための方法論を明らかにする．ここで，日本語・英語の利用可能な語彙資源を数多く組み合わせて用いることにより，収集した語彙ギャップに関するデータを定量的に分析する計算論的アプローチをとる．

(2) 異言語の語彙概念の対応付け

日本語，英語の具体的な意味的語彙資源として，WordNet，および，EDR 電子化辞書を用い，語彙概念の対応付け手法を開発する．手法としては，対訳資源を用いて対応先の候補を求め，これを目的言語の知識を用いて順位付けるという手法をとる．特に，このために有用な手がかり情報を明らかにする．

(3) 異言語の対応付けの情報構造モデル

近年では，言語資源を Web 上でサービス化する動きが活発化していることを踏まえ，各国語の意味的な語彙資源がサービス化されている環境において，主に(2)により検討される手法により漸進的に対応付けられていく異言語の語彙概念により形成される多言語の意味ネットワークを表現する枠組みを，語彙資源に関する国際標準である LMF (Lexical Markup Framework) なども参考にしながら提案する．

４．研究成果

(1) 語彙化を判定する基準・手法

まず，語彙ギャップ事例の収集を行った．語彙ギャップの事例収集のためには，まず対訳辞書に記載されている訳語の語義を特定する「語義対応付け」が必要となる．このために対訳辞書における訳語表現と単言語辞書における語義定義文との異言語表現間の類似度を数量化するマッチング手法を実装した．ここで，具体的な語彙資源として，単言語・概念辞書(日本語: Lexeed, EDR 電子化辞書, 英語: WordNet, EDR 電子化辞書, LDOCE)，対訳辞書(EDR 電子化辞書，EDICT)を用いた．

次に，訳語情報として与えられている表現が語彙化されているかを判定する検討を進めた．より具体的には，原言語(日本語)における見出し語に対する制約条件に記述のパターン，訳語(英語)の表記のバリエーションに関する分析を進めた．さらに，対訳辞書で与えられている英訳語が英語の意味辞書 WordNet において，どのように語義分けされ，説明・定義されているかを分析した．また，このような言語資源をオントロジーとして見なす観点から既存研究，および，今後の研究の方向性を解説論文（雑誌論文-4)としてまとめた．

(2) 異言語の語彙概念の対応付け

今回開発した対応付けの方法論は，与えられたある言語の語彙概念に対応しうる別の言語における語彙概念を探索するためのものである．そこで，対訳資源を用いて対応先の候補となる目的言語における語彙概念の候補集合をまず求め，この中で各候補の語彙概念に対して意味的な関連度を計算する方法をとる(図 1)．

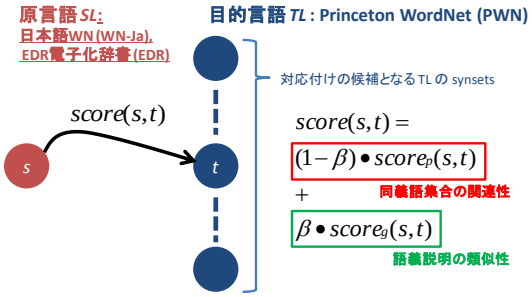

図 1: 異言語の語彙概念の対応付けの方式

図 1 の枠組みにおいて，関連度を計算するための手がかりとしては様々な情報が考えられるが，まず，原言語・目的言語における語彙概念を規定する同義語集合の言語横断的な関連度のみを用いる手法の評価を行い，次に，これに加えて語彙概念(語義)を説明するテキストの言語横断的な類似性を加味する有効性の評価を行った．

① 同義語集合の有効性評価: 本手法では，原言語の語彙概念の同義語集合を構成する各語を複数の対訳資源を用いて目的言語に翻訳し，目的言語の語彙概念の同義語集合との関連度を計量化する．ここで，必要となることは，以下に不適切な翻訳な多義を抑制すること，さらに，もともとが原言語において同義語である各語の翻訳語集合の一貫性がどの程度保持されているかを見積もることである．前者の課題の解決のためには，目的言語における語義タグ付きコーパス(より具体的には Princeton Annotated Gloss

Corpus）が有用であることを示し，さらに，後者に関しては，適切な重み付けを行うための関数を実験的に導いた．このようにして開発した方式を実際の語彙資源を用いて評価した結果，日本語 WordNet から選出した 4,690 個の語彙概念をクエリとして用いる対応付けの再現実験においては，上位 5 件の候補に対して約 73%の成功率を得ることができ，また，WordNet と異なる成り立ち・構造を持つ EDR 電子化辞書の約 200 個の概念を語彙概念と見立ててクエリとして用いる対応付けの発見実験においても約 69%の成功率を得ることができた．この結果を国際会議（雑誌論文-1），および，国内学会研究会（学会発表-2）において発表した．

② 語義説明テキストの有効性評価：上記の結果は同義語集合の言語横断的関連度のみに着目するものであるため，さらに，語義説明テキストの類似性や，それぞれの意味体系における局所的な構造の類似性を加味することにより，精度を向上させることが期待できる．そこで，同義語集合の情報に加え，語義説明テキストが手がかり情報としていかに有効に利用できるかを検討した．より具体的には，原言語の語義説明テキストを複数の機械翻訳システムにより目的言語に翻訳し，目的言語の語彙概念に付与されている語義説明テキストとの類似性を評価し，この類似度を①の手法による関連度と重み付き統合することの有効性を評価した．図 2，図 3 はそれぞれ，再現実験，発見実験における結果を示す．横軸は語義説明テキストの類似度に対する重みであり，縦軸は精度である．

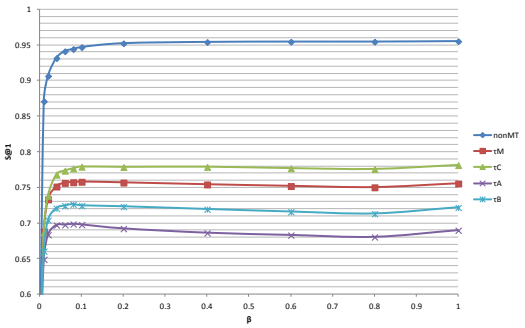

図 2：再現実験における語義説明テキスト類似度の有効性評価

再現実験（図 2）においては，容易に予想できるように語義説明テキストの有効性が明らかであるが，おおむね β=0.2 程度で同等の精度が得られていることから，原言語と目的言語の語彙資源の類似性が高い場合は，この程度の重みが適切であるといえる．また，機械翻訳の冗長は有用に働いている．

発見実験（図 3）においては，再現実験と全く異なる傾向が示された．すなわち，適切な重み（おおむね $\beta \leqq 0.1$）を設定することにより精度は若干向上するが，その制御は容易ではない．また，複数の機械翻訳を用いることの冗長性の効果は必ずしも明らかではなく，そもそもの出自の異なる異言語の語彙資源の対応付けにおいては，語義説明テキストの類似性を有効に活かすことは困難な課題であることが明らかとなった．これにより，説明テキストの性質を言語的に検討し，類似性が有用に働くケースを予測する手法の開発が必要であることが明らかとなった．この結果を国内学会年次大会において発表した（学会発表-1）．

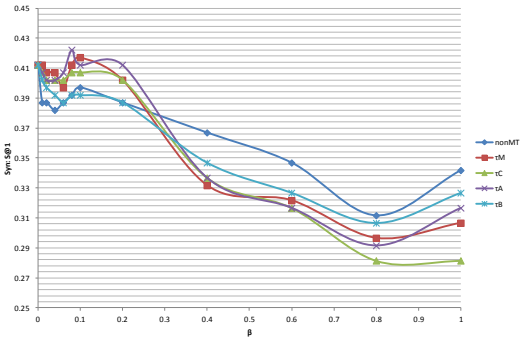

図 3：発見実験における語義説明テキストの類似性の評価

(3) 異言語の対応付けの情報構造モデル

Web 上の言語サービス基盤においては，Web サービス化された言語資源を Web サービス技術によって組み合わせることにより，新たな言語資源を仮想的に実現できる．さらに，言語グリッドのような協調的な言語サービス環境においては，ユーザの利用を通して漸進的に成長していくような新たなタイプの言語資源を考えることができる．そこで，単言語の意味概念辞書と対訳辞書を組み合わせることによる仮想的な複合辞書へのアクセスサービスを想定し，言語間で対応する語彙概念をユーザによるアクセス要求を契機とし機会主義的に対応付けることにより，動的かつ漸進的に成長していく多言語の意味ネットワーク・語彙的オントロジーを表現するための枠組みを提案し，さらに，これらの意対応関係を二次的な言語資源とするための要件について検討した．

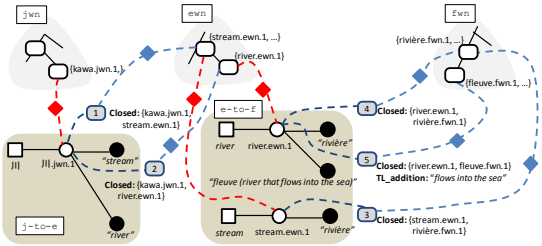

図 4：多言語間の語義（語彙概念）の対応付け

図 4 は，漸進的に得られる「川」（日本語），

″river/stream″（英語），″rivière /fleuve″（フランス語）の語義(語彙概念)の間の対応付けに対する表現を模式的に表したものであるが，より具体的な提案として，語彙資源のモデル化・表現に関する国際標準である LMF (Lexical Markup Framework)に基づく場合の拡張要素を明らかにした．

上記は，語彙ギャップが存在しない場合の例であるが，語彙ギャップが存在する場合は，語彙化されない言語側での言語表現に対するオブジェクトを生成し，これと語彙化された概念の間の対応付けを適切に表現することが必要となる．さらには，語彙化が困難であるために，単語より複雑な言語表現を有している場合においては，その構成要素との関係が規定できる場合があり，このような状況を適切に表現できることが求められる．図 5 はそのようなケースを示す．

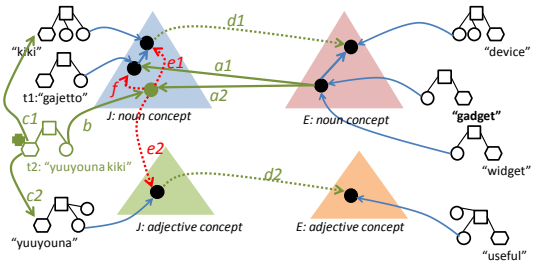

図 5: 語彙ギャップの構成的対応付け

この例では，英語の″gadget″に対して，日本語で「有用な機器」という説明的翻訳，および，「ガジェット」という音訳が与えられていて，さらに″gadget″の語義説明テキストとして″useful device″というフレーズが与えられている場合を想定している．この場合，それぞれの語義説明テキスト要素の間に useful/有用な，道具/device という対応を抽出することができるので，語彙ギャップが存在する場合においても，より多くの言語間の対応関係を抽出することが可能となる．

以上の結果を国際会議(雑誌論文-2,3)，および，国内学会研究会(学会発表-3)において発表した．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

1. Yoshihiko Hayashi. Computing Cross-Lingual Synonym Set Similarity by Using Princeton Annotated Corpus. Proceedings of the 6-th International Global WordNet Conference (GWC2012), 査読有, pp.134-141 (2012)
2. Yoshihiko Hayashi. Direct and Indirect Linking of Lexical Objects for Evolving Lexical Linked Data. Proceedings of the Second Workshop on the Multilingual Semantic Web (MSW2011), 査読有, pp.62-67 (2011)
3. Yoshihiko Hayashi. A Representation Framework for Cross-Lingual/Interlingual Lexical Semantic Correspondences. Proceedings of the 9th International Conference on Computational Semantics (IWCS2011), 査読有, pp.155-164 (2011)
4. 林 良彦. 言語的オントロジーの構築と展開. 人工知能学会誌, 依頼原稿・閲読有, Vol.25, No.3, pp.335-344 (2010)

〔学会発表〕（計 3 件）

1. 林 良彦. 異言語の語彙概念の対応付けのための手がかり情報の有効性評価. 言語処理学会第 18 回年次大会, 発表論文集 pp.617-620, 2012.3.15, 広島市立大学.
2. 林 良彦. Princeton Annotated Corpus を用いた異言語の語彙概念の対応付け. 電子情報通信学会・思考と言語研究会, Vol.111, No.428, pp.29-34, 2012.2.4, 機械振興会館.
3. 林 良彦. 協調的な言語サービス基盤上における複合辞書アクセスサービスの検討. 電子情報通信学会・人工知能と知識処理研究会, Vol.110, No.428, pp.1-6, 2011.2.28, 関西学院大学.

〔図書〕（計 2 件）

1. 林 良彦. 言語的オントロジーの構築と展開. 来村徳伸（編著）. オントロジーの普及と応用. オーム社, pp.67-90 (2012)
2. 林 良彦. 言語処理・機械翻訳. 白井克彦（編著）. 音声言語処理の潮流. コロナ社, p.205-238 (2010)

６．研究組織

(1)研究代表者

林　良彦（HAYASHI YOSHIHIKO）
大阪大学・大学院言語文化研究科・教授
研究者番号：80379156

(2)研究分担者

なし

(3)連携研究者

永田　昌明（NAGATA MASAAKI）
日本電信電話株式会社・コミュニケーション科学基礎研究所・主幹研究員
研究者番号：10418551

(研究計画立案時～H21 年度に参画)