DBpediaのカテゴリ情報を利用したIs–aリンク構築支援の検討

Is–a Relationship Construction Support Using Category Information in DBpedia

山元 悠太 [1,3] *　古崎 晃司 [2] †　駒谷 和範 [3] ‡
Yuta Yamamoto[1,3]　Kouji Kozaki[2]　Kazunori Komatani[3]

[1] 大阪大学大学院 工学研究科
[1] Graduate School of Engineering, Osaka University
[2] 大阪電気通信大学 情報通信工学部
[2] Faculty of Information and Communication Engineering, Osaka Electro-communication University
[3] 大阪大学 産業科学研究所
[3] The Institute of Scientific and Industrial Research, Osaka University

Abstract: 本研究では，オントロジーを利用したドキュメント分析への適用を想定したオントロジー拡張の手法について述べる．ドキュメントとオントロジーの関連付けが弱い場合，分析に必要な情報が十分に取得できない可能性がある．そこで本研究では，Web 上の構造化データである Linked Open Data (LOD) を利用した，ドキュメントに応じたオントロジーの拡張の支援を行うシステムを目指す．本研究では，オントロジー拡張の 1 ステップである Is–a リンク構築について，先行研究に基づき，LOD の 1 つである DBpedia のカテゴリ情報を用いた手法を提案した．実際に拡張が行われたオントロジーを利用して，先行研究の手法との性能比較実験を行った結果，構築できる Is–a リンクの数と正解率が共に上昇し，提案手法の有効性を確認した．

1　導入

近年，様々な分野において，ドキュメント (文書) 分析への需要が高まっている．ドキュメント分析は，ドキュメントの文法構造などを利用したグラフ化や，他のデータとの関連付けによって行われ，話題抽出や，それに応じた自動分類，情報検索といった用途がある．このタスクにおいて，ドキュメント中で表現された情報や，それらの意味関係を取得する方法として活用が期待されているのが**オントロジー**である．

オントロジーとは，システム上で人間の持つ知識を体系化したデータである．オントロジーは，ノードと，ノードの間を繋ぐエッジで構成され，このうちノードは**概念**，エッジは**関係**，**リンク**とも呼ばれる．これらの要素は人手で追加されていくことが多い．しかし，作業者の知識や思考などの要因によって差異が生じやすく，また人件費や時間といったコスト面での問題もある．本研究では，この問題を解決するために，オントロジーを構成する概念・関係を自動的に追加することで拡張を行う方法について考える．

オントロジーを使ったドキュメント分析の課題として，ドキュメントとオントロジーの関連付けがある．ドキュメント分析を行う際には，ドキュメント中の単語とオントロジーの概念を関連づける処理が重要である．この時にオントロジーが持つ概念 (**既存概念**) の数が十分でなければ，ドキュメントに対する情報取得も十分に行えないため，分析を満足に行えないことがある．

この課題の解決策として，本研究では，ドキュメントを利用したオントロジー拡張支援システムの構築を目指す．これは，関連するドキュメントに応じて，必要な概念を自動的にオントロジーへ追加することで，ドキュメント中でオントロジーと関連付けられる単語を増やすことを目指すものである．

このシステムでは，オントロジーとそのドメインに関するドキュメントを入力として，オントロジーに新し

*yamamoto@ei.sanken.osaka-u.ac.jp
†kozaki@osakac.ac.jp
‡komatani@sanken.osaka-u.ac.jp

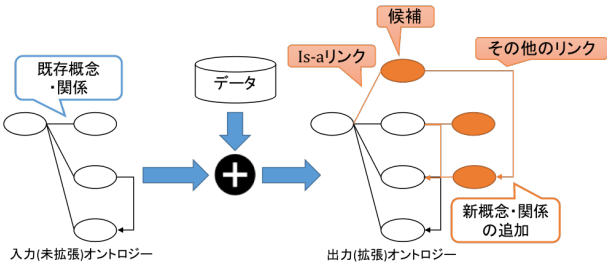

図 1: オントロジーとその拡張のイメージ

く概念として追加する単語 (**追加概念**) と，その**追加先概念** (追加概念との間にリンクを構築する既存概念) の候補を出力する．各追加概念に対する追加先概念を決める際には，Web 上のデータである Linked Open Data (LOD) を利用する．そして，出力の 1 つである追加位置は，1 つのみに断定するのではなく，適切さに応じたランキングの形式とする．

2　オントロジーの拡張

2.1　拡張のステップ

オントロジーの拡張は，主に以下の 3 つの作業ステップから成る (図 1)．

1. 概念として追加する単語の選定 (候補選定)
2. Is–a リンクの構築
3. その他のリンクの構築

1 つ目のステップである候補選定では，オントロジーに新しく追加する概念を決める．これは文書 [1, 2, 3] の他，Wikidata[1]や DBpedia[2]といった LOD，および他のオントロジーなどの構造化データを使って [4, 5, 6] 決められる．ここで選ばれた全ての候補が追加されるとは限らない．

2 つめのステップにある Is–a リンクは，オントロジーで基本的なリンクとされる上位/下位関係のことである．上位概念は下位概念をより一般化した概念であり，下位概念は上位概念の情報を原則的にすべて継承する．このステップでは，前のステップで決定した概念候補を，追加先であるオントロジーの適切な概念の子として追加する．また，拡張のステップにおいて参照する情報として，Linked Open Data (LOD) という Web 上のデータを利用する．

最後のステップでは，オントロジーに Is–a リンクで追加した概念に対して，Is–a 以外の関係で他の概念とのリンクを構築する．この関係には，ある概念の一部であることを示す Part–of 関係，付随する情報を示す Attribute–of 関係，同じものであることを示す SameAs 関係，などがある．こうしたリンクの追加により，概念に対する多面的な情報を表現することが可能となる．

本研究では，ドキュメント中の単語を使って候補選定することを想定している．しかし現段階では主に次の Is–a リンク構築に取り組んでいるため，以降では取り扱わない．

2.2　関連研究

オントロジーの拡張，およびその各ステップに関する研究は多数行われてきた．本節では，候補選定，Is–a リンク構築に関する研究を取り上げる．

候補選定に関係する研究では，松尾ら [7] や Litvak らの研究 [8] の研究がある．松尾らは，ドキュメントから重要な語句を抽出する方法として，文書全体の単語出現頻度と，ある単語と他の単語の共起頻度の違いを利用した．これは，重要語句は他の重要な語句と共起しやすいという仮説に基づき，分布頻度の差についてランク付けを行うことで，重要な語句が上位に来ることを確かめた．Litvak ら [8] は，文書から抽出した重要語句の各候補について，他の文書への参照数，および他の文書からの被参照数を利用した HITS[9] という尺度によってその重要度を算出する手法を考案した．これにより，コーパスなどを用いた教師あり学習による重要語句選定と同等の性能を，少数の文書からの教師なし学習で発揮できることを示した．

リンク構築のステップに関する研究では，Wille[1] によって提唱された形式概念分析 (Formal Concept Analysis; FCA) が知られている．これはドキュメント中の文構造を利用してリンクを構築する手法であり，Text2Onto[2] や OntoGain[3] など，ドキュメントだけを利用したオントロジー構築手法でよく採用されている．しかし FCA は，LOD などの背景知識などを利用しないため，性能的な面で課題が残っている．DBpedia などの構造化データを利用した研究では，Klink[4] や Klink2[5]，多田らの研究 [6] などが挙げられる．多田らの研究は本研究のベースとして扱ったため，詳細は 3.2 節に示す．

3　Is–a リンク構築

3.1　DBpedia

本研究では，LOD の 1 つである DBpedia の情報を利用した Is–a リンク構築を行う．DBpedia は，オンライン百科事典である Wikipedia[3]から情報を自動抽出し

[1] https://www.wikidata.org
[2] http://ja.dbpedia.org/
[3] https://ja.wikipedia.org

図 2: DBpedia における検索結果の一例

てグラフ構造化することで構築された LOD である．なお，DBpedia をはじめとする LOD では，グラフのノードに当たる情報を**エンティティ**と呼ぶ．DBpedia では，Wikipedia の各言語版における項目名をそれぞれ対応するエンティティのラベルとして，他の項目との関係などをリンクとして持つ (リンクのラベルは意味ごとに個別に定義されている)．

一例として，DBpedia における「地球温暖化」の概念に関するリンクの一部を図 2 に示す．この例における各リンクは，`dcterms:subject` はその項目が属するカテゴリ，`rdfs:comment` は概要テキスト，`rdfs:label` は項目名，`owl:sameAs` は他言語版の DBpedia などにおける対応項目を表している．

本研究で主に用いるのは，これらリンクのうち，項目の所属カテゴリの情報である `dcterms:subject`，および，カテゴリ間での上位/下位関係である `skos:broader` の 2 種類である．また，一部のエンティティは，特定の別のエンティティを指す項目であることを示す情報として，`dbpedia-owl:wikiPageRedirects`(リダイレクト) というリンクを持つ場合がある．この場合には，詳細な情報を持つリダイレクト先エンティティと同一のものとして扱う．

オントロジー中の概念と DBpedia 中の項目の関連付けは，それぞれのラベルが完全一致するもののみを採用した．今後の研究では，DBpedia Spotlight[10] をはじめとする Wikification[11] を利用して，より高度な関連付けを行いたいと考えている．

3.2　ベースライン手法

3.2.1　カテゴリの共通性を利用した Is–a リンク構築

本研究で Is–a リンク構築のベースとするのは，多田らの研究 [6] で提案された，DBpedia のカテゴリ情報階層を利用した手法である (図 3)．この手法では，追加概念候補と，既にオントロジー中にある概念に対して DBpedia 中の対応する項目を取得し，そのカテゴリ情報を利用することで Is–a リンクを構築する．この手法は「上位概念が同じなら，下位概念のカテゴリの系列にも共通性がある」という考えに基づいて提案されたものである．詳細な手順を以下に示す．なお，追加先候補である概念は，予めユーザがオントロジーで定義されている既存概念から複数選択することで決めておく (オントロジー中の全概念としても良い)．

1. オントロジーから，追加先候補である既存概念 u_i の下位概念の集合 $\overline{\boldsymbol{U}}_i$ を取得する
2. 各下位概念 $\overline{u}_{ij} \in \overline{\boldsymbol{U}}_i$ に対応する DBpedia のエンティティを取得し，その所属カテゴリを最大で 10 段辿り[4]，カテゴリ集合 $\boldsymbol{C}_i$ を得る
   - 各エンティティに対する所属カテゴリ (1 段目) は `dcterms:subject` のリンクから取得する
   - 2 段目 (カテゴリの上位カテゴリ) 以降は `skos:broader` のリンクから取得する
3. カテゴリ $c_{ik} \in \boldsymbol{C}_i$ それぞれについて確からしさ $confidence(c_{ik}|u_i)$ を算出する (式 1)
   - $under(\overline{\boldsymbol{U}}_i|c_{ik})$ を下位概念 $\overline{\boldsymbol{U}}_i$ のうちカテゴリ c_{ik} に属するものの数とする

$$confidence(c_{ik}|\overline{\boldsymbol{U}}_i) = \frac{under(\overline{\boldsymbol{U}}_i|c_{ik})}{|\overline{\boldsymbol{U}}_i|} \quad (1)$$

   - カテゴリ c_{ik} が複数のカテゴリ集合 $\boldsymbol{C}_i$ で出現した場合，全ての $\boldsymbol{C}_i$ から除外する (確からしさを算出しない)
   - 追加先候補 u_i とカテゴリ c_{ik} が確からしさの値を持って 1 対 1 で対応するようになる

[4] DBpedia ではカテゴリに上位–下位の階層構造がある

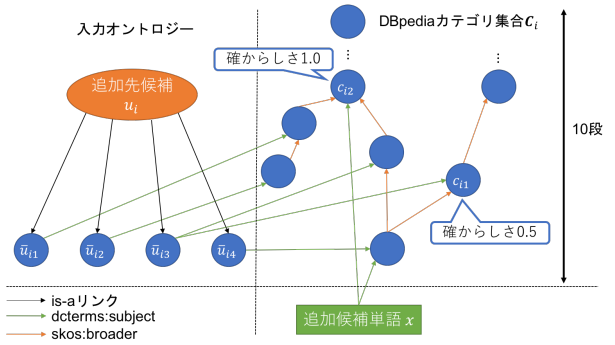

図 3: 多田ら [6] の Is–a リンク構築手法

4. 追加概念候補 x の所属カテゴリ集合 $\boldsymbol{C}_x$ を取得する

5. カテゴリ $c_{xl} \in \boldsymbol{C}_x$ が $\boldsymbol{C}_i$ の中にあれば，その確からしさが最も高い c_{xl} を探す (なければ u_i は上位概念にならない)

6. 最も高い確からしさを持つカテゴリ c_{xl} に対応する u_i を上位概念とし，x との間に Is-a リンクを構築する

3.2.2　ベースライン手法の課題

ベースライン手法では，指標として DBpedia におけるカテゴリ情報の共通度合いを利用していた．しかしここには，以下の 2 つの課題があった．1 つは，所属カテゴリを 10 段も辿る必要があるのか，という点である．カテゴリ階層を辿っていくと，上層になるほど元の概念とは関係がなさそうなものになりやすくなる．例えば,「地球温暖化」という項目の所属カテゴリを辿ると，1 段目は「環境問題」であるが，2 段目は「社会問題」，3 段目は「社会倫理学」となっていき，5 段目では「行動」となる．多田らは，こうしたカテゴリ階層で何段目にあるかによって重みは付与せず，全て同列に扱っていた．また，追加先候補の下位概念に対してはカテゴリを 10 段辿るのに対し，追加概念候補に対しては 1 段しか辿らないため，共通するカテゴリが見つからない可能性も考えられる．

もう 1 つの課題は，複数の系列で出現するカテゴリを考慮しなかったことである．この手法では，カテゴリの確からしさについて，追加先候補と 1 対 1 対応になるようにしていた．この時，複数の追加先候補に対する系列，つまり，辿った 10 段分のカテゴリ階層に，複数の系列で出現するカテゴリがあれば，確からしさを算出しないようにしていた．追加先候補の概念とカテゴリを 1 対 1 対応にするためには，複数の概念と関連がある (曖昧性がある) カテゴリは考慮するべきではない，としていたためである．しかしこれによって，追加先候補が増えると重複するカテゴリの数も増加してしまい，1 対 1 対応を取ることが難しくなる可能性があった．

3.3　カテゴリの近さを利用した Is–a リンク構築

前節の課題を踏まえ，本研究では，追加候補と既存概念が属するカテゴリ階層を辿った際の経路の距離を利用する (図 4)．この Is–a リンク構築は，以下の手順で行う．

1. オントロジーから，追加先候補である既存概念 u_i の下位概念の集合 $\overline{\boldsymbol{U}}_i$ を取得する

2. 追加概念候補 x と各下位概念 $\overline{u}_{ij} \in \overline{\boldsymbol{U}}_i$ の所属カテゴリを，共通するカテゴリが見つかるまで 1 段ずつ辿っていく

   - x から辿った段数と $\overline{u}_{ij}$ から辿った段数の和: $hop(x, \overline{u}_{ij})$

3. 各 u_i に対するカテゴリ階層上の平均経路長 $avghop(u_i)$ を算出し (式 2)，これが最小となる u_i と x との間に Is–a リンクを構築する

   - h を各追加先候補から追加概念候補までのホップ数の合計とする

$$avghop(u_i) = \frac{\sum_{\overline{u}_{ij} \in \overline{\boldsymbol{U}}_i} hop(x, \overline{u}_{ij})}{|\overline{\boldsymbol{U}}_i|} \quad (2)$$

この手法によって，ベースライン手法の課題の解決を図った．1 つ目の課題であるカテゴリを 10 段遡ることについては，カテゴリを辿って共通するものが見つかった時点で処理を終了できるため，必要な段数を減らすことができる．2 つめの課題であった複数系列に出現するカテゴリについても，この手法では各カテゴリを経路の通過地点として見なすに留めるため，曖昧性を考慮する必要はなくなる．

4　性能比較実験

本章では，2 つの分野のオントロジーを使って行った，Is–a リンク構築手法の性能を比較する実験について述べる．比較を行ったのは，多田ら [6] の手法 (ベースライン) と，その改良案として提案した 3.3 節の手法である．

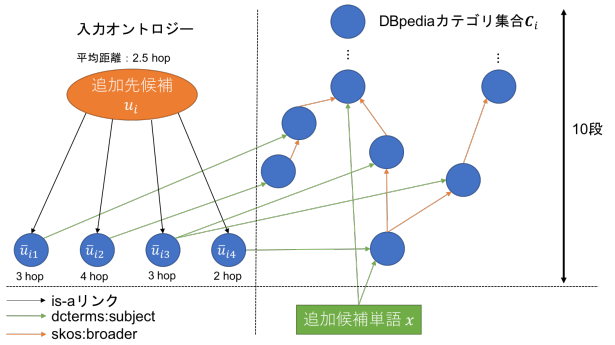

図 4: Is–a リンク構築法の改善

4.1　サステイナビリティ分野での実験

この実験は，提案手法の性能を試験的に確かめるために行った．本実験の条件では正解データが存在しなかったため，結果を手作業で評価することで性能を確かめた．正解データがある条件下で行った実験については 4.2 節で述べる．

4.1.1　サステイナビリティオントロジー

サステイナビリティオントロジーは，主に環境対策に関する情報を定義したオントロジーである．このオントロジーの概念には，環境問題やそれに対する対策・目標の他，対策の評価方法，およびこれらに関係する物質，変化，行為などを持つ．この実験で使用したサステイナビリティオントロジーは，概念を 4,527 個持つバージョンである．また，全ての概念・リンクは，専門家の監修のもとで手作業で追加されたものである．

このオントロジーに含まれる 4,527 個の概念のうち，DBpedia にラベル完全一致のエンティティが存在したものは 1,178 個 (26.1%) であった．概念の中には，助詞や助動詞などを含む形式のもの (資源に対する制限，経済学的な回復力，など) が多く含まれており，これらは対応する項目を取得することが難しいため，エンティティのカバー率が少なくなってしまったと思われる．

4.1.2　実験方法

ドキュメントから追加概念の候補単語 (追加候補単語) を選定した後，Is–a リンク構築を行った．追加候補単語を選定するドキュメントには，総合地球環境学研究所から発表されている研究要覧[5]の 2018 年度版を使用した．このテキストを形態素解析器である MeCab[6] にかけ，名詞と判定された形態素 4,431 語を追加候補として選定した．そして各単語に対して，ベースライン手法と本研究の提案手法の 2 通りの方法で上位概念を決定した．

本実験では，オントロジーへの追加候補である単語に対する上位概念の候補を，人工物，物質，状態，自然構造物，製品，の 5 つの概念のみに限定した．この理由は，リンクの構築結果に対して，どの概念が上位概念としてふさわしいか，という正解データが存在しないため，手作業で正解/不正解を判別して評価する必要があり，その簡易化を図るためである．これら 5 概念は，その下位概念が一定数存在し，かつ Wikipedia に項目があるものが多いためカテゴリ情報が取得しやすいものを選定した．

[5] http://www.chikyu.ac.jp/publicity/publications/brochure/

[6] https://taku910.github.io/mecab/ 分かち書き辞書には ipadic-NEologd(https://github.com/neologd/mecab-ipadic-neologd) を使用

表 1: サステイナビリティ分野でのリンク構築実験結果
() 内の数字は追加単語の総数 4,431 個に対する割合を表す

手法	全構築リンク数	正解数	Precision
多田	292 (0.07)	92	0.32
提案	1,558 (0.35)	590	0.38

4.1.3　結果

実験結果を表 1 に示す．この結果より，本研究の提案手法によって，多田の手法よりも構築できるリンクの数，そのうちの正解率 (Precision) が共に上昇したことがわかる．なお，本実験における正解率は，構築結果である上位概念が，それぞれ実際に上位概念とした場合に相応しい (正解) か相応しくない (不正解) かを手作業で評価したものである．

4.2　生物規範工学分野での実験

前節で使用したサステイナビリティオントロジーは，概念の追加に対する正解データがなかった．それに対し本実験では，実際に大規模な拡張が行われたオントロジーを用いて，提案手法のより定性的な評価を行った．

4.2.1　生物規範工学オントロジー

生物規範工学は，生物の身体構造などを分析し，工学的に利用することを目的とする学問のことである．よって，生物規範工学オントロジーの概念には，生物の分類や，その身体構造，生態，性質に関するもの，及びこれらに関係する物質，変化などが含まれる．

生物規範工学オントロジー [12] には，拡張前と拡張後のバージョンが存在する．この拡張は，書籍に記さ

れた表現を基に概念の名前やリンクを決定して，人手で追加することで行われた．拡張前のバージョンでは概念が 1,366 個，拡張後のバージョンでは概念が 1,615 個含まれている．単純な概念数の差では 249 個増加しているが，拡張によって拡張前のバージョンから削除された概念が 263 個あるため，実際に拡張によって追加された概念の数は 512 個である．

オントロジーの拡張によって追加された 512 個の概念のうち，DBpedia にラベル完全一致のエンティティが存在したものは 144 個 (28.1%) である．本実験では，DBpedia に対応するエンティティが存在する追加概念 144 個に対して，上位概念を決定して Is–a リンクを構築することで性能評価を行った．

4.2.2　実験方法

本実験の手順について述べる．Is–a リンク構築の手順については 4.1 節と同様であるが，構築を行う対象は 4.2.1 項で述べた単語 144 個とした．

この実験では，実際に拡張を行ったデータを正解データとして用いて評価を行った．本実験で採用した評価基準として，各追加概念に対して，拡張データで上位概念とされている概念 (正解) と，Is–a リンク構築の結果として出力された概念 (構築結果) の距離を利用した．ここで言う距離とは，オントロジー上で概念間の Is–a リンクを辿った時に必要となるホップ (直接的にリンクを持つ概念間の移動) の数である．以降この距離のことを概念間距離と呼ぶ．

オントロジー上では，ある 2 概念の概念間距離が短いほど，意味が近しくなる．そのため，概念間距離に閾値を設け，Is–a リンクの構築に成功した，と見なす範囲を設定した．概念間距離の閾値を 3 とした場合の例を，図 5 に示す．正解である 1 つの概念に対して，そこから概念間距離が閾値以下である他の概念が構築結果となった場合，リンク構築に成功したと見なす．図 5 の場合には，点線で囲った範囲の概念すべてが成功の範囲になる．つまり，正解である既存概念が 1 つだけであるのに対し，リンク構築に成功したと見なす既存概念は複数存在する．

4.2.3　結果

本実験で．構築できた Is–a リンクの総数は，多田の手法で 20 個，提案手法で 100 個となった．実験に用いた追加概念の数が 144 個のため，多田の手法では 14%，提案手法では 69%がカバーできたことになる．

本実験では，4.1 節の実験のように上位概念を 1 つだけ出力するのではなく，上位概念候補のランキングを出力した．このランキングは，ベースライン手法の場合は確からしさが大きい順にカテゴリに対応する上位概念候補を，提案手法では，平均経路長が短い順に上位概念候補を並べることで作成した．この出力は，閾値を変えた際，成功する数がどのように変化するかを調べる条件を揃えるため，最大で 10 位までとした．以下，ランキングの内容を利用して行った 2 種類の評価について述べる．

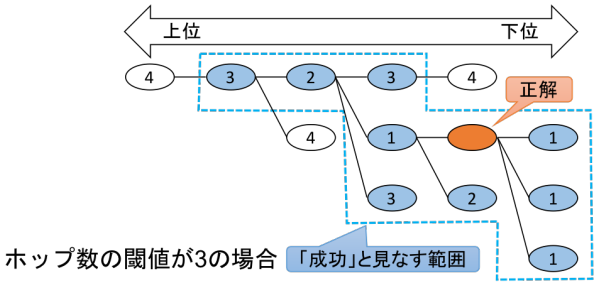

図 5: リンク構築の成功判定の例
各概念ノード内に示した数字は，正解概念からの概念間距離を表す．

表 2: 生物規範工学分野でのリンク構築実験結果：構築成功数

閾値	構築に成功したリンク数	
	ベースライン	提案手法
0	1	25
1	1	31
2	3	46
3	**4**	**58**
4	4	72
5	6	79

1 つ目は，構築できたリンクのうち成功した数，すなわち，各手法で決定された上位概念が正解の上位概念に近かったものについての評価である．この評価では，ランキング中の順位は考慮せず，成功と見なす概念が上位 10 位に入っていた数をカウントした．ホップ数の閾値を変化させた際のカウントを，表 2 に示す．

この表より，いずれの閾値においても，多田の手法に比べて本稿の提案手法によってリンク構築が成功する割合が増加したことがわかる．成功と判定する閾値を 3 とした場合で 58%，閾値 0，つまり構築結果と正解が一致していなければならない場合で 25%の割合で成功する結果となった．

2 つ目は，先の結果における順位についての評価である．4.2.2 項で述べたように，この評価では，リンク構築に成功したと見なす既存概念が複数存在する．そこで，出力されたランキングのうち，成功と見なす既存概念が最高で何位に順位付けられるかを確認した．評価は，それぞれの上位概念候補のうち，ホップ数が閾値以下である概念の最高ランクの平均を取ることで行っ

表 3: 生物規範工学分野でのリンク構築実験結果：最高順位平均

閾値	リンク構築に成功したと見なす概念の最高順位の平均	
	ベースライン	提案手法
0	1.0	3.6
1	1.0	3.2
2	1.0	3.1
3	**1.0**	**2.6**
4	1.0	2.4
5	1.0	2.0

た．その結果を，表 3 に示す．

この表より，提案手法によって，閾値 3 の場合，平均で 3 位以内には正解に意味の近い概念が出現することが分かった．閾値を 0 とした場合でも，平均で 4 位以内には正解に意味の近い概念がランクインした．一方，多田らの手法の場合，元々順位付けを目的にはしていなかったため，構築結果を 2 つ以上出力することはできなかった．

4.3　考察

4.2 節の実験結果より得られた考察について述べる．Is–a リンク構築の結果を見て判断した結果，概念間距離が 2 以内であれば十分に意味が近いと言える例が多く見受けられた．例えば，カメムシ目，ハエ目，といった目階級の昆虫の分類概念は，昆虫という概念を経由して 2 ホップで移動でき，その下の科階級には 3 ホップで移動できる．他には節足動物という概念に 2 ホップ，動物という概念に 3 ホップで移動できるが，4 ホップ以上では動物の他の下位概念である魚や哺乳類など，分類が大きく異なるものが出てくる．そのため，閾値 3 を成功判定の基準とした場合に得られた知見を述べる (基準となる実験結果は表 2，3 に太字で示す)．

本研究の提案手法によって，上位概念のランキングを作成すると，そのうち 58%で，Is–a リンク構築が成功したと見なせる概念が入ることが分かった (閾値を 3 と置いた場合)．また，このランキングのうち成功である概念の最高順位の平均は閾値が 0 の場合で 3.6 位，閾値が 3 の場合で 2.6 位であった．

これらを踏まえると，オントロジーに追加したい単語に対して，

- DBpedia に対応するエンティティが存在する
- カテゴリ情報が定義されている

という条件を満たしていれば，

- 上位概念のランキングが作成できる
- このランキングのうち 58%に，正解の上位概念，もしくはそれに意味の近い概念が含まれる
- 上記の概念はランキングの 3 位以内にランクインする

という結果となることが分かる．この結果は，完全に自動でオントロジー拡張を行う[7]には不十分だが，ユーザに対して上位概念の候補を 3 個から 5 個程度提示し，その中から最終決定をしてリンクを構築する，というシステムとしては利用できる可能性がある．こうした性能に関しては，十分な性能であると言えるような定量的指標が存在しない．しかし，サステナビリティオントロジーの構築・拡張を進めている環境ドメインの専門家からは「このような提案が行えれば，オントロジー拡充が容易になるであろう」とのコメントを得ている．サステイナビリティオントロジーの利用については，前述のように評価が難しいため，人手で正解データを作成するなどの方法による応用用途として進めていく予定である．

多田の手法を使った場合，本研究の提案手法と比較して，構築できた Is–a リンクの数はかなり少ないという結果となった．この理由は，3.2.2 節で述べたように，複数のカテゴリ階層で出現するカテゴリは考慮しないため，追加先候補が 1,000 個以上に増えた影響であったと考えられる．また，本研究の提案手法では追加単語・追加先候補の両方について最大 10 段までカテゴリを辿るのに対し，多田の手法では追加単語に対して 1 段しかカテゴリを辿らなかったことも影響したと思われる．

5　結論

本研究では，オントロジーの拡張，そのうち Is–a リンク構築の段階について，DBpedia のカテゴリ情報を利用したリンク構築について論じた．先行研究で提案されていた，カテゴリ情報の共通性を利用した手法の課題を改善し，カテゴリ情報の経路の近さを利用した手法を考案した．2 つの手法の性能を比較する実験を行い，提案手法が有効であることを確認した．しかし，現段階の性能では，選定した追加概念候補に対して完全自動で Is–a リンクを構築するには不十分である．構築性能の向上のため，追加概念候補と DBpedia の項目との関連付けの改良を目指す他，カテゴリ情報以外の DBpedia の情報[8]をまだ利用していないため，今後取り入れていきたい．

[7]追加したい単語に対して，システム上でユーザの意思に関わらず Is–a リンクを自動で構築することを指す．

[8]Wikipedia で各項目の基本情報などを表す Infobox や，項目について簡潔に述べた概要文など

また，Is–a リンク構築に強く関係するステップである追加概念の候補選定についても考えていきたい．本研究では，ドキュメントから有用な単語，および複合語を選定して新しい概念として追加することを想定している．DBpedia Spotlight[10] をはじめとする．文脈情報などを利用してテキスト中の単語を Wikipedia の項目や LOD のエンティティと関連付ける Wikification[11] の手法を導入することで，関連付けを強化し，この目標を達成したい．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (B)17H01789 の補助を受けて実施された．

参考文献

[1] Rudolf Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. In *Proceedings of the 7th International Conference on Formal Concept Analysis*, ICFCA '09, pp. 314–339. Springer-Verlag, 2009.

[2] Philipp Cimiano and Johanna Völker. Text2onto. In Andrés Montoyo, Rafael Muńoz, and Elisabeth Métais, editors, *Natural Language Processing and Information Systems*, pp. 227–238. Springer Berlin Heidelberg, 2005.

[3] Euthymios Drymonas, Kalliopi Zervanou, and Euripides G. M. Petrakis. Unsupervised ontology acquisition from plain texts: The ontogain system. In *Proceedings of the Natural Language Processing and Information Systems, and 15th International Conference on Applications of Natural Language to Information Systems*, NLDB'10, pp. 277–287. Springer-Verlag, 2010.

[4] Francesco Osborne and Enrico Motta. Mining semantic relations between research areas. In *The Semantic Web – ISWC 2012*, pp. 410–426. Springer Berlin Heidelberg, 2012.

[5] Francesco Osborne and Enrico Motta. Klink-2: Integrating multiple web sources to generate semantic topic networks. In *Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366*, pp. 408–424, 2015.

[6] 多田恭平, 古崎晃司, 來村徳信, 溝口理一郎, 駒谷和範. 概念間の関係に注目した専門文書解析と LOD 技術によるバイオミメティクス・オントロジーの大規模化の試み. 人工知能学会全国大会論文集, Vol. JSAI2015, pp. 1–4, 2015.

[7] Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, Vol. 13, pp. 157–169, 2003.

[8] Marina Litvak and Mark Last. Graph-based keyword extraction for single-document summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, pp. 17–24. Association for Computational Linguistics, 2008.

[9] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, Vol. 46, No. 5, pp. 604–632, 1999.

[10] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pp. 1–8. ACM, 2011.

[11] Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pp. 233–242. ACM, 2007.

[12] 古崎晃司, 來村徳信, 溝口理一郎. 生物規範工学オントロジーと Linked Data に基づくキーワード探索. 人工知能学会論文誌, Vol. 31, No. 1, pp. 1–12, 2016.