Web 検索を用いた関連のある概念間の関係抽出手法

白川　真澄[†]　　中山浩太郎[††]　　荒牧　英治[††]

原　　隆浩[†]　　西尾章治郎[†]

† 大阪大学大学院情報科学研究科マルチメディア工学専攻
〒 565-0871 大阪府吹田市山田丘 1-5
†† 東京大学知の構造化センター
〒 113-8656 東京都文京区本郷 7-3-1
E-mail: †{shirakawa.masumi,hara,nishio}@ist.osaka-u.ac.jp, ††nakayama@cks.u-tokyo.ac.jp,
†††eiji.aramaki@gmail.com

あらまし　固有名詞や専門用語，及びこれら概念間の関係を網羅した大規模オントロジの構築は，意味中心の Web を実現する上で重要な基盤技術である．最近の研究では，Wikipedia における概念の網羅性を活用した大規模オントロジの自動構築に注目が集まっているが，is-a 関係や a-part-of 関係など，形式的な関係のみを対象としており，概念間の関係の網羅性は低い．そこで本研究では，Wikipedia マイニングによって抽出した情報に Web 全体の情報を組み合わせることで，概念および概念間の関係の種類を網羅した大規模オントロジの自動構築を目指す．本稿では，大規模オントロジの自動構築の核となる関係抽出の手法として，Wikipedia から抽出した連想シソーラスを利用して関係の深い概念ペアを取り出し，Web 検索によって概念間の関係を抽出する手法を提案する．

キーワード　関係抽出，自然言語処理，シソーラス，Wikipedia

A Relation Extraction Method between Related Concepts using Web Search

Masumi SHIRAKAWA[†], Kotaro NAKAYAMA[††], Eiji ARAMAKI[††],

Takahiro HARA[†], and Shojiro NISHIO[†]

† Department of Multimedia Engineering, Graduate School of Information Science and Technology,
Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
†† The Center for Knowledge Structuring, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
E-mail: †{shirakawa.masumi,hara,nishio}@ist.osaka-u.ac.jp, ††nakayama@cks.u-tokyo.ac.jp,
†††eiji.aramaki@gmail.com

Abstract　Construction of a huge scale ontology covering many named entities, domain specific terms and relations among these concepts is one of the essential technologies in the next generation Web based on semantics. In this work, we aim at automated ontology construction with a wide coverage of concepts and these relations by combining information on the Web with Wikipedia. In this paper, we propose a relation extraction method which gets pairs of co-related concepts from an association thesaurus extracted from Wikipedia and extracts relations by using Web search.

Key words　relation extraction, natural language processing (NLP), thesaurus, Wikipedia

1. 研究背景

近年，意味中心の Web の実現に注目が集まっており，これまで様々な研究が行われてきた．その中でも，世の中のあらゆる概念間の関係を定義した大規模オントロジの構築は，意味情報を考慮した情報検索をはじめ，多種多様なアプリケーションの基盤技術として必要とされている．これまで行われてきたオントロジ構築に関する研究は，特定の用途に特化したタスク依存のオントロジか，あるいは一般的な概念のみを対象とする上位オントロジの構築を目的とする研究がほとんどであった．前者は主に，タスクの要求に合わせたオントロジを人手で構築する際の手助けとなるツールの開発であり，これまでに数多くの実用的なツールが開発・利用されている [1]．一方，本研究とも関係のある後者は，対象とする範囲が世界全体であるようなオントロジの構築に関する研究である [2]．実際に構築・公開されている上位オントロジは，OpenCyc(注1)や WordNet [3]，EDR 電子化辞書(注2)など，一般的な概念を対象として，is-a 関係や a-part-of 関係などの形式的な関係のみを定義したオントロジが多い．しかし，情報検索をはじめとする様々なアプリケーションの要求に応えるためには，固有名詞や専門用語，およびこれら概念間の様々な関係を定義した大規模オントロジが必要である．最近では，大規模な Web 百科事典である Wikipedia を用いて，固有名詞や専門用語を網羅したオントロジの自動構築を試みる研究が数多く行われている．しかし，これらの研究では，Wikipedia の概念の網羅性を主張する一方，概念間の関係の種類についてはあまり重要視していないのが現状である．

そこで本研究では，Wikipedia から抽出した情報に，Web 全体の情報を組み合わせることで，以下の二点双方に着目した大規模オントロジの自動構築を目指す．

- 固有名詞および専門用語の網羅性
- 概念間の関係の種類の豊富さ

具体的には，固有名詞および専門用語を網羅するため，Wikipedia から抽出した大規模連想シソーラス [4], [5] を利用して関係の深い概念ペアを取り出す．これは，固有名詞や専門用語を網羅するというだけでなく，関係を定義すべき概念ペアを絞り込めるというメリットがある．その後，豊富な関係を抽出するため，Web 全体を対象とした検索を行う．概念ペアにそれぞれ格助詞を付与したクエリを生成して Web 検索を行い，レスポンスに含まれる Web ページを，自然言語処理を用いて解析することで，概念間の関係を抽出する．

2. 関連研究

大規模オントロジの自動構築に関する研究はこれまで数多く行われてきたが，最近では，Wikipedia を対象とした Web マイニングにより，大規模オントロジの自動構築を試みる研究が注目を集めている．Wikipedia は，Wiki をベースにした大規模 Web 百科事典であり，誰でも Web ブラウザを通じて記事内容を変更できることが大きな特徴である．そのため，幅広い分野について，一般的な概念から新しい概念に至るまで記事が網羅されており，記事（概念）数は，最も多い英語版で 300 万記事，日本語版で 60 万記事である（2009 年 12 月時点）．また，Wikipedia は，記事の網羅性や即時性だけでなく，密で多様なリンク構造，質の高いリンクテキスト，URL による語彙の一意性など，知識抽出のコーパスとして有利な特徴を数多く持っている [6]．このような特徴を持つ Wikipedia を解析し，大規模オントロジを構築した例として，DBpedia [7] と YAGO [8], [9] が挙げられる．DBpedia では，Wikipedia の記事に定義されているインフォボックスと呼ばれる構造化データに着目し，インフォボックスから情報を抽出して RDF (Resource Description Framework) に変換するという簡潔な手法により，多言語にわたる大規模なオントロジの自動構築に成功している．YAGO では，英文法に特化した自然言語処理技術を利用し，一般語を対象としたオントロジである WordNet のクラスに Wikipedia のカテゴリをマッピングすることで，WordNet を拡張した大規模なオントロジを自動構築している．また，日本語の大規模オントロジ構築の試みとしては，隅田らの研究 [10] や桜井らの研究 [11] が挙げられる．隅田らの研究では，Wikipedia のカテゴリ構造やリスト構造などから is-a 関係（上位下位関係）の候補を抽出し，機械学習手法である SVM (Support Vector Machine) を用いてフィルタリングを行うことで，高精度なオントロジを構築している．桜井らの研究では，日本語版 Wikipedia の様々な半構造化データを用いてオントロジの自動構築を行い，日本語版 Wikipedia のオントロジ構築への適用可能性を示している．これらの研究では，セマンティック Web の実現を目指していることから，is-a 関係や a-part-of 関係など，定義・形式化が容易な関係のみを対象としたオントロジ構築を目標としている．そのため，概念の関係の種類については，あまり網羅されていないのが現状である．

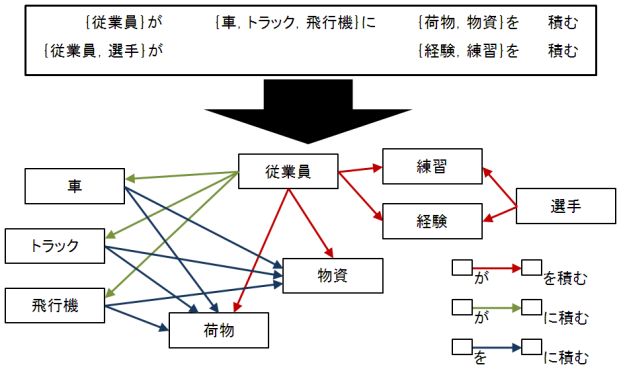

図 1　格フレーム辞書からオントロジの再構築
Fig. 1　Ontology reconstruction from Case Frames

一方，概念間の様々な関係について定義している辞書として，格フレーム辞書 [12], [13] がある．格フレーム辞書では，膨大な数の文書を解析し，名詞・格助詞・動詞の共起情報を基に名詞のクラスタリングを行うことで，名詞間の関係を動詞によって表現している．格フレーム辞書は，図 1 のように，名詞を概念，

(注1)：http://www.opencyc.org/
(注2)：http://www2.nict.go.jp/r/r312/EDR/J_index.html

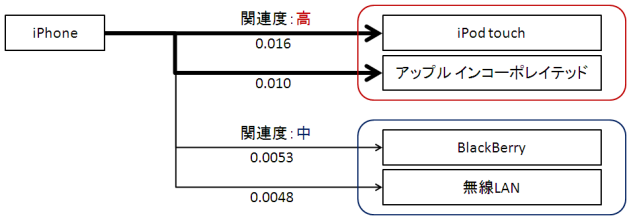

図 2　Wikipedia シソーラスの例
Fig. 2　An example of Wikipedia Thesaurus

動詞と格助詞の組み合わせを関係として再構築することで，一種のオントロジとみなせる．このようにして再構築されたオントロジは，概念間の関係を様々な動詞によって柔軟に表現できる．しかし，固有名詞や専門用語については，一般名詞と比べて共起情報が少ないため，あまり定義されていない．

オントロジあるいは辞書を構築するという目的とは異なるが，固有名詞間の様々な関係を抽出する手法として，数原らの手法 [14] が挙げられる．数原らは，話題語によるブログでの検索結果のスニペットを解析対象とし，動詞と格助詞を軸に話題語を繋げることで動作関係を抽出している．ただし，数原らの手法は，定常的な関係（例：貿易を行う）よりも一時的な関係（例：貿易を開始する）の抽出を目的とした手法である．本研究で提案する手法では，定常的な関係と一時的な関係を同等に扱っている点で異なる．また，数原らの手法では，箇条書きやリストなど，関係抽出に適さない Web ページを多く取得してしまう場合があると述べられている．一方，大島らの手法 [15] では，抽出対象を同位関係のみに限定しているが，クエリ生成の時点で並列助詞を付与することで，同位関係抽出に適した Web ページを効率的に絞り込んでいる．

3. 提案手法

本研究では，関連のある概念ペアが与えられたときに，それら概念間の関係を表す動詞を，Web 検索を用いて抽出する手法を提案する．

3.1 関連のある概念ペアの抽出

次節以降で説明する Web 検索を用いた概念間の関係抽出手法は，関連のある概念ペアを入力として関係を抽出する手法である．そのため，オントロジを自動構築するに当たって，先に関連のある概念ペアを用意する必要がある．そこで本研究では，Wikipedia シソーラス [4], [5] を利用する．Wikipedia シソーラスは，Wikipedia で定義されている概念間の関連度を定義した辞書である．例えば，図 2 のように,「iPhone」と「iPod touch」は関連度が 0.016（相対的に高い）,「iPhone」と「BlackBerry」は関連度が 0.0053（相対的に中程度），といった情報が定義されている．また，関連のない概念ペアについては，関連度は定義されていない．Wikipedia シソーラスを用いることで，何らかの関係が存在する概念ペアを推測できるため，Web 検索によって関係を抽出すべき概念ペアを絞り込める．また，Wikipedia シソーラスは,「アンパンマン」と「やなせたかし」や,「ダイクストラ法」と「エドガー・ダイクストラ」など，固有名詞や専門用語についても関連度が定義されており，既存のオントロジでは定義されていないような，専門性の高い概念ペアについても関係を抽出できると考えられる．

表 1　概念「ソフトバンク」のリンクテキスト一覧と信頼度 $R_{\text{ソフトバンク}}$
Table 1　A list of link texts of “Softbank” and reliability

リンクテキスト	使用回数	$R_{\text{ソフトバンク}}$
ソフトバンク	371	1
日本ソフトバンク	17	0.479
SoftBank	4	0.234
ゲームバンク	4	0.234
ソフトバンク（株）	3	0.186
SOFTBANK	3	0.186
Softbank	3	0.186
ソフトバンクグループ	2	0.117
ソフトバンク株式会社	2	0.117
ソフトバンクショップ	2	0.117
ソフトバンクテレコム株式会社	1	0
ソフトバンクパブリッシング	1	0
同グループ	1	0

3.2 Web 検索を用いた概念間の関係抽出

関連のある概念ペアが与えられたとき，Web 検索を用いてこの概念間の関係（動詞）を抽出する手法を提案する．提案手法では，概念ペアを入力すると，出力として概念ペアの関係が順位（重要度）付きで得られる．ここで得られる関係は，概念ペアに付属する格助詞と動詞の組み合わせによる関係である．また，重要度は，頻度情報やシノニム（同義語）情報を基にして算出される．例えば,「孫正義」と「ソフトバンク」の概念ペアを入力すると,「孫正義がソフトバンクを創業する」「孫正義がソフトバンクを経営する」といった関係が，重要度の高い順に出力される．提案手法では，概念ペアにそれぞれ格助詞を付与したクエリを生成して Web 検索を行う．その後，レスポンスの上位に含まれる Web ページのテキストから，概念ペアが出現する文を抽出し，係り受け解析・形態素解析を行うことで動詞を取得する．以下，各プロセスについて詳述する．

3.2.1 各概念のシノニムと信頼度の取得

自然言語処理を用いた概念間の関係抽出では，表記の揺れや省略形など，概念を表す表現が複数存在するため，これらの表現をうまく認識できない場合，網羅性や精度が低下する．そこで，Wikipedia のリンクテキストから抽出したシノニム（同義語）情報を利用する [16]．Wikipedia では，各概念の記事に他の概念が出現すると，その概念の記事に対してリンクが張られる．このとき，リンクテキストは，リンク先の概念を端的に表した表現が使用されるため，このリンクテキストを各概念のシノニムとして利用できる．例えば，Wikipedia の全記事に対して，概念「ソフトバンク」へのリンクテキストを全て取得すると，表 1 のような情報が得られる．表 1 のリンクテキストを全て，概念「ソフトバンク」のシノニムとして利用できるが，リンクテキストとしての使用回数が多いほど，そのリンクテキストがシノニムとして適していることが分かる．そこで，概念 c へのリンクテキスト集合を L_c として，リンクテキスト $l_i \in L_c$ の使用回数 $M_c(l_i)$ を基に，l_i のシノニムとしての信頼度 $R_c(l_i)$

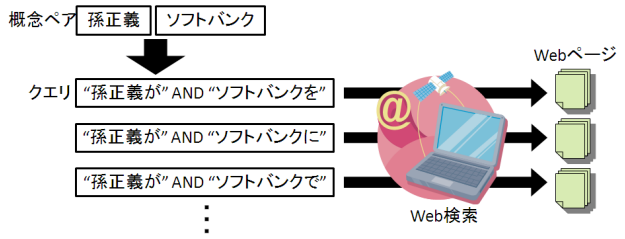

図 3　格助詞を付与した Web 検索クエリ生成の例
Fig. 3　An example of query generation with case particles

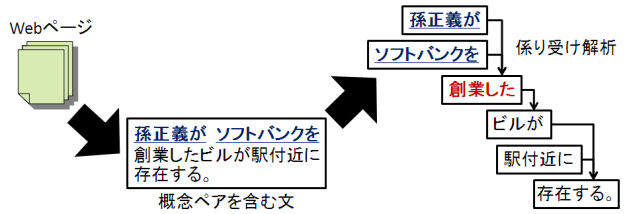

図 4　レスポンス解析による関係抽出の例
Fig. 4　An example of relation extraction by analyzing responses

を以下の式により定義する．

$$R_c(l_i) = \frac{\ln(M_c(l_i))}{\ln(\max(M_c(l_k))}, \ l_k \in L_c \tag{1}$$

信頼度 $R_c(l_i)$ は，リンクテキスト l_i が概念 c を表現するのにどの程度適しているかを表しており，1 に近いほど，その概念のシノニムとして適している．なお，使用回数 $M_c(l_i)$ が 1 を超えるリンクテキスト l_i が存在しない場合，Wikipedia の URL として用いられている概念名のみをシノニムとする．信頼度 R_c は，3.2.3 節で述べる重要度 I の計算に用いる．また，信頼度 R_c の最も高いシノニム，つまりリンクテキストとして最も使用されている表現を，その概念を表すラベル名とし，次節でクエリを生成する際に用いる．

3.2.2　格助詞を付与したクエリによる Web 検索

提案手法では，クエリ生成の時点で概念ペアの各概念に格助詞を付与し，Web 検索を行う．数原らの研究 [14] では，格助詞を付与せずに Web 検索を行っているが，箇条書きやリスト，題名などの部分にクエリがヒットし，格助詞が出現する文がほとんど取得できず，関係を抽出できない場合があると述べられている．そこで，クエリ生成の時点で格助詞を付与することで，概念ペアの関係を表す動詞を抽出するのに適した Web ページの絞り込みを行う．例えば，「孫正義」と「ソフトバンク」の概念ペアについて，格助詞「が」と「を」をそれぞれ付与して AND 検索を行うと，レスポンスには「孫正義が」と「ソフトバンクを」を確実に含む Web ページがヒットする．これらの Web ページには，箇条書きやリスト中ではなく，文章中に概念ペアが記述されている可能性が高い．したがって，これらの Web ページから「孫正義」と「ソフトバンク」の関係を表す動詞が得られる可能性は，格助詞を付与しないクエリで検索した場合と比べて高くなると考えられる．

具体的には，図 3 のように，それぞれの格助詞の組み合わせについてクエリを生成する．その後，各クエリについて Web 検索エンジンで AND 検索を行い，各々についてレスポンスの上位 N 件の Web ページと，レスポンスの総ヒット数 H を取得する．レスポンスの総ヒット数 H は，3.2.3 節で述べる重要度 I の計算に用いる．

3.2.3　レスポンス解析による関係抽出

各クエリによってそれぞれレスポンスの上位 N 件の Web ページを取得した後，句点やピリオド，改行などの区切り文字によって文単位に分割し，その中から各概念が両方含まれる文のみを抽出する．このとき，概念が文中に含まれているかどうかの判定に，3.2.1 節で説明したシノニム情報を用いる．例えば，文中に「SoftBank」という表現が出現した場合，表 1 を参照し，この表現が概念「ソフトバンク」を意味している可能性があると判断する．

その後，概念ペアの両方が含まれている文に対し，CaboCha [17] による係り受け解析および MeCab [18] による形態素解析を行い，概念ペアの両方が係っている動詞を抽出する．具体的には，自立語の動詞（例：争う，投げる，積む）とサ変接続の名詞（例：販売，生産，デビュー）を，関係を表す動詞として抽出する．なお，「する」は自立語の動詞であるが，関係を表す動詞として意味を成さないため，抽出対象から除外する．抽出した動詞と，そのとき概念ペアに付属していた格助詞の組み合わせを一つの関係とみなし，関係ごとに出現回数をカウントする．抽出した動詞を v，動詞 v に係る概念ペア c_1，c_2 のシノニムをそれぞれ l_i，l_j，概念ペアに付属する格助詞をそれぞれ p_1，p_2 とし，この関係を抽出したときの個別の出現回数 $F(p_1, p_2, v)$ を以下の式により算出する．

$$F(p_1, p_2, v) = \frac{R_{c_1}(l_i) + R_{c_2}(l_j)}{2} \tag{2}$$

この F は，実質的にこの関係が何回分出現したとみなすか，という出現回数を意味している．シノニム l_i，l_j が共にそれぞれの概念を表すのに最も代表的な表現であれば 1 回分出現したとみなされる．そうでなければ，シノニムが意味する概念が，c_1 や c_2 ではない可能性があるとみなし，1 回に満たない回数分（例：0.5 回分など）しか出現していないものとする．そして，全ての文を解析した後，それぞれの関係の重要度 $I(p_1, p_2, v)$ を以下の式により算出する．

$$I(p_1, p_2, v) = \max\left(\frac{H(p_1, p_2)}{N}, 1\right) \cdot \sum F(p_1, p_2, v) \tag{3}$$

なお，$H(p_1, p_2)$ は格助詞 p_1，p_2 をそれぞれ付与したクエリによるレスポンスのヒット数，N は Web ページの取得時の制限数である．式 (3) は，Web 全体でこの関係がどの程度存在しているかを推測している．この重要度 I の高い順に，関係の順位付けを行う．

4.　評 価 実 験

提案した概念間の関係抽出手法について評価実験を行った．クエリに格助詞を付与する手法の有効性を検証するため，格助詞を付与しないクエリによって Web ページを取得する手法と比較を行った．また，提案手法によって実際に抽出した関係を

俯瞰し，既存のオントロジで定義されていないような関係が抽出できているかを検証した．

4.1 評価環境

評価方法として，関連のある概念ペアに対して，各手法を用いて関係（動詞）を抽出し，人手により評価を行った．具体的には，Wikipedia シソーラスからランダムに選出した 306 の固有名詞同士の概念ペアについて，格助詞「は」「が」「を」「に」「で」をそれぞれ付与したクエリ（全 20 通り）を生成して Web 検索を行った．ここでは，レスポンス数が上位 5 つまでの格助詞の組み合わせについてのみ，Web ページを 50 件ずつ取得（全 250 ページ）した．各手法によって得られた関係のうち，重要度の高い上位 5 つの関係を，人手により正解 (H)，部分正解 (A)，不正解 (B) に判定した．判定の基準として，概念ペアと動詞，格助詞のみで関係が明確に理解できるものを正解，補足情報無しでは関係が明確に理解できないもの（例：「岡山市が岡山大学に新設」何を新設したかが重要な情報である）や場合によっては正しくない関係（例：「清水エスパルスがジュビロ磐田に勝利」敗北することもある）を部分正解，その他を不正解とした．また，正しく文を解析できている場合は，その関係が事実か否かに関係なく正解とした．つまり，Web 上で噂として話題に挙がっている関係も正解に含める．評価指標には，数原らの研究 [14] で用いられている以下の三つの指標を用いた．

（1） 最上位の正解率 (CFR)

最上位に順位付けされた関係の正解率である．

（2） Mean Reciprocal Rank (MRR)

最も上位である正しい関係の順位の逆数をとり，平均をとる．正しい関係が存在しない場合は 0 とする．

（3） Discounted Cumulative Gain (DCG)

Jarvelin らによって提案された評価指標である [19]．情報検索の分野において，多段階の適合性評価を行うための指標であり，以下の式によって表される．

$$dcg(i) = \begin{cases} g(1) & \text{if } i = 1 \\ dcg(i-1) + \dfrac{g(i)}{log_c(i)} & \text{otherwise} \end{cases} \tag{4}$$

$$g(i) = \begin{cases} h & \text{if } r(i) \in H \\ a & \text{if } r(i) \in A \\ b & \text{if } r(i) \in B \end{cases} \tag{5}$$

なお，$r(i)$ は i 番目に順位付けされた関係であり，$g(i)$ は $r(i)$ の正解の度合いに基づく適合度，$dcg(i)$ は上位 i 番目までに順位付けされた関係の適合度の累積である．各パラメータは文献 [14], [20] と同様に $(h, a, b) = (3, 2, 0)$，$c = 2$ とした．

なお，全く関係を抽出できなかった概念ペアについては，上記三つの評価指標において 0 とみなす場合と，評価指標の算出の対象外とする場合に分けた．また，CFR と MRR では，部分正解を含む場合と含まない場合に分け，DCG では上位 5 位までを対象とする場合と上位 3 位までを対象とする場合に分けた．

4.2 評価結果

評価結果を表 2，表 3 にそれぞれ示す．また，評価実験における解析データを表 4 に記載する．

表 2 評価結果（全概念ペア）
Table 2 Evaluation (all concept pairs)

評価指標	比較手法	提案手法
CFR（部分正解含む）	0.141	0.258
CFR（部分正解除く）	0.0817	0.160
MRR（部分正解含む）	0.145	0.272
MRR（部分正解除く）	0.0866	0.181
DCG（5 位まで）	0.547	1.27
DCG（3 位まで）	0.519	1.14

表 3 評価結果（関係抽出した概念ペア）
Table 3 Evaluation (concept pairs relation extracted between)

評価指標	比較手法（49 ペア）	提案手法（101 ペア）
CFR（部分正解含む）	0.878	0.782
CFR（部分正解除く）	0.510	0.485
MRR（部分正解含む）	0.908	0.825
MRR（部分正解除く）	0.541	0.549
DCG（5 位まで）	3.42	3.86
DCG（3 位まで）	3.24	3.44

表 4 評価実験における様々な解析データ
Table 4 Analysis data in the experiment

解析データ	比較手法	提案手法
解析の対象となった Web ページ数	72473	37019
概念ペアの両方が含まれる文の数	1650	11197
関係を抽出した回数	140	953
関係を抽出した割合	0.0848	0.0851
関係抽出した概念ペア数	49	101
抽出した関係数	93	263
概念ペア当たり関係数（全ペア）	0.304	0.859
概念ペア当たり関係数（関係抽出したペア）	1.90	2.60

まず，CFR についてみると，全概念ペア（306 ペア）を対象とした場合，部分正解を含む含まないに関わらず，提案手法が比較手法と比べて 2 倍近くの評価値を達成している．これは，抽出した関係の中で優先度が 1 位である関係についてみたとき，提案手法の方が比較手法よりも正しい関係を 2 倍近く抽出できていることを意味している．また，最低一つ以上関係を抽出できた概念ペアのみを対象としたとき，部分正解を含む場合は，提案手法のほうが評価値がやや低くなっているが，部分正解を除く場合は，ほぼ同程度となっている．つまり，提案手法は，比較手法と同程度の抽出精度で 2 倍近く正しい関係を抽出できている．

MRR についてみると，全概念ペアを対象とした場合，CFR と同様，部分正解を含む含まないに関わらず，提案手法の評価値が比較手法と比べて 2 倍近く大きい．また，関係を抽出した概念ペアのみについては，部分正解を含む場合は CFR のときと同様に比較手法がやや勝っているが，部分正解を除く場合はほぼ同等となっている．これは，一つの概念ペアあたり多くの関係を抽出しているほうが値が高くなる，という MRR の性質によるものである．MRR では，最も上位である正しい関係の

表 5　出力結果の例
Table 5　An example of extracted relations

概念 A	概念 B	抽出された関係
愛・地球博	長久手町	A が B で開催
Berryz 工房	夏焼雅	A に B が加入
NEWS	山下智久	A を B が脱退
ZARD	坂井泉水	A は B が唄う
サザンオールスターズ	桑田佳祐	A を B が代表
ザルツブルク	モーツァルト	A で B が作曲
ノースウエスト航空	デルタ航空	A が B に吸収
フィアット	ランチア	A が B を生産
シューベルト	ベートーヴェン	A は B を尊敬
プロレスリング・ノア	三沢光晴	A を B が旗揚げ
一関市	東山町	A に B は合併
京都大学	東京大学	A は B に次ぐ
京都産業大学	京都成安高等学校	A に B を移管
古田敦也	ヤクルト	A が B に入団
藤岡弘、	仮面ライダー	A が B に変身
香取慎吾	SMAP	A が B を休む

順位の逆数をとるが，得られた関係の中に不正解のものしか存在しない場合，値が 0 となってしまう．比較手法では，関係を抽出できた概念ペアについて，一つの概念ペアあたり 1.90 個の関係しか抽出できていないのに対し，提案手法では 2.60 個の関係を抽出している．つまり，提案手法では，比較手法では抽出できなかった概念ペアについて関係を抽出できているだけでなく，同じ概念ペアで比べても，比較手法よりも多く関係を抽出できている．

DCG に関しては，どの場合においても提案手法が優っている．これは，DCG では 1 位から 5 位（あるいは 3 位）まで，抽出された全ての関係が値に影響し，且つ，評価値が加算方式であるためである．つまり，抽出精度に関わらず，より多く正しい関係を抽出しているほど，DCG は高くなる．そのため，関係を抽出した数が多い提案手法のほうが値が高い．

解析データをみると，提案手法よりも比較手法のほうが，解析の対象となった Web ページの数が多い．これは，格助詞を付与したクエリによる Web 検索では，レスポンス数が 50 件に満たない場合が多いためである．しかし，概念ペアの両方が含まれる文の数をみると，提案手法のほうが圧倒的に多くなっている．同様に，関係を抽出した回数も，概念ペアの両方が含まれる文の数に比例している．これは，Web 検索クエリに格助詞を付与することによって，関係抽出に適した Web ページを絞り込めていることを表している．

4.3　出力結果

評価実験において，実際に提案手法によって抽出された出力結果の一例を表 5 に示す．表 5 は，それぞれの概念ペアについて最も優先度が高かった関係を表しているが，既存のオントロジでは定義されていないような様々な関係が抽出できていることが分かる．例えば,「シューベルトはベートーヴェンを尊敬」といった人物間の内面的な関係や,「藤岡弘、が仮面ライダーに変身」といった役者と役の関係を，動詞によって詳らかに表現している．また,「プロレスリング・ノアを三沢光晴が旗揚げ」「京都産業大学に京都成安高等学校を移管」といった関係では，類似した意味を持つ動詞（例：旗揚げ，創設，創業）の中でも特に相応しい動詞（実際に使われている動詞）が抽出できている.「京都大学は東京大学に次ぐ」といった一位と二位の順位関係も，二者間の関係として抽出できており，動詞による関係の柔軟性を表している．興味深いのは，2010 年 1 月初めに香取慎吾がインフルエンザにより SMAP を休んだという事実が，関係抽出を行った 2010 年 1 月末において，関係として抽出されていることである．新しい話題やニュースに関する関係は，定常的な関係よりも優先して抽出される傾向にあるが，このような関係も，既存のオントロジにはない関係であり，有益な情報であると考えられる．この中で,「NEWS を山下智久が脱退」という関係は，2010 年 1 月末においては，噂として広まっているだけで事実ではないが，アプリケーションへの応用を考えた場合，それ自体が事実でなくても，十分に有益な情報と成り得る．

表 6　誤って抽出された関係の例
Table 6　An example of extracted wrong relations

概念 A	概念 B	抽出された関係
MBS ラジオ	毎日放送	A が B を設立
オークランド	ウェリントン	A が B に変身
ドラえもん	藤子・F・不二雄	A は B が輝く
ミハイル・ゴルバチョフ	ボリス・エリツィン	A は B が使用
メルボルン	シドニー	A は B で退役
今治市	今治駅	A で B に一泊
吉永小百合	浜田光夫	A が B に言う
国際放映	東宝	A が B は設立
奈良テレビ	テレビ和歌山	A は B にリレー
日野自動車	いすゞ自動車	A が B は販売
松田優作	探偵物語	A が B で住む

一方，表 6 は，誤って抽出された関係の一例を表している．誤った関係が抽出されてしまった原因はいくつか存在するが，主な原因としては，誤ったシノニム情報，誤った係り受け解析，三者間以上の関係のうち重要でない二者間の関係の抽出，が考えられる．一つ目の例として,「MBS ラジオが毎日放送を設立」という関係の場合,「MBS ラジオ」のシノニムの中に「MBS」,「毎日放送」のシノニムの中に「MBS テレビ」という誤った情報がそれぞれ存在しており，実際は「MBS」が「MBS テレビ」を設立しているのに対し,「MBS ラジオ」が「毎日放送」を設立したことになっている．二つ目の，誤った係り受け解析は，誤った関係を抽出する最も大きな原因であった.「ドラえもんは輝く」「いすゞ自動車は販売」といった通常の文に，もう一つの概念が同じ動詞に誤って係ることで，意味の通らない関係が抽出されている．三つ目は，例えば「松田優作が探偵物語で住む」という関係では,「住む」という動詞は，実際はその直後に出現する「事務所」というキーワードが重要であるが，そのキーワードが失われており,「住む」という動詞単体では概念間の関係として意味を成さないため，誤った関係となっている．

関係を抽出できなかった概念ペアは，306 ペア中，205 ペアであった．これらの概念ペアの傾向として，is-a 関係や a-part-of

関係を持つ概念ペア，作品や番組，楽曲などが片方に含まれる概念ペアが大部分を占めていた．このうち，is-a 関係や a-part-of 関係については，既存手法により取得できるため，提案手法によって新たに関係を抽出する必要はない．しかし，後者の関係は，様々な関係パターンが考えられ，既存手法では抽出できない．これは，もう片方の概念と動詞によって明示的に関係が記述されているような文があまり存在しないことが原因である．また，「8 時だョ!全員集合」や「クリスマスキャロルの頃には」などのように，概念のラベル名が文である場合が多いことも，係り受け解析の精度に影響するため，関係を抽出できない一因であると考えられる．

5. まとめと今後の課題

本稿では，Wikipedia マイニングによって抽出した情報に，Web 全体の情報を組み合わせることで，概念および概念間の関係の種類を網羅した大規模オントロジを自動構築するための関係抽出手法を提案した．具体的には，Wikipedia から抽出した連想シソーラス（Wikipedia シソーラス [4], [5]）を利用して関係の深い概念ペアを取り出し，格助詞を付与したクエリにより Web 検索を行い，概念間の関係を抽出する手法を提案した．

評価実験の結果，クエリ生成時に格助詞を付与することにより，2 倍程度の正しい関係を，同程度の精度で抽出できることを確認した．また，提案手法によって抽出できた関係は，既存のオントロジでは定義されていないような特殊な関係が多く，関係の種類を網羅した大規模オントロジを構築する上で有効な手法であるといえる．抽出された誤った関係については，いくつか共通した原因がみられ，解決方法を検討する必要がある．

今後の課題として，網羅性および精度向上のための手法の拡張を行う必要がある．例えば，全格助詞の組み合わせを用いることや，動詞以外の関係の抽出を行うことで，関係の種類の網羅性を向上する手法を考案する予定である．

また，もうひとつの課題として，実際に大規模オントロジを構築することが挙げられる．大規模オントロジを構築し，公開することは，オントロジに関係する分野の発展において非常に重要である．例えば，オントロジ構築に携わる研究者は，公開されたオントロジを土台として，より優れたオントロジを構築できる可能性がある．また，オントロジを基盤リソースとしたアプリケーションの研究開発に携わる技術者は，公開されたオントロジを用いて有益なサービスを生み出せる可能性がある．このような観点から，提案した手法を用いて，実際に大規模オントロジの構築・公開を目指す．

謝辞 本研究の一部は，科学研究費補助金基盤研究 C(20500093)，科学研究費補助金基盤研究 B(21300032)，およびマイクロソフト産学連携研究機構 CORE 連携研究プロジェクトの助成によるものである．ここに記して謝意を表す．

文　献

[1] 古崎晃司，溝口理一郎，“オントロジー構築ツールの現状，” 人工知能学会誌，vol.20，no.6，pp.707–714，Nov. 2005.
[2] 武田英明，“上位オントロジー，” 人工知能学会誌，vol.19，no.2，pp.172–186，Mar. 2004.
[3] G.A. Miller, “WordNet: A Lexical Database for English,” Communications of the ACM (CACM), vol.38, no.11, pp.39–41, Nov. 1995.
[4] 中山浩太郎，原隆浩，西尾章治郎，“Wikipedia マイニングによるシソーラス辞書の構築手法，” 情報処理学会論文誌，vol.47，no.10，pp.2917–2928，Oct. 2006.
[5] K. Nakayama, T. Hara, and S. Nishio, “Wikipedia Mining for An Association Web Thesaurus Construction,” Proceedings of International Conference on Web Information Systems Engineering (WISE), pp.322–334, Dec. 2007.
[6] 中山浩太郎，原隆浩，西尾章治郎，“人工知能研究の新しいフロンティア: Wikipedia，” 人工知能学会誌，vol.22，no.5，pp.693–701，Sept. 2007.
[7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z.G. Ives, “DBpedia: A Nucleus for a Web of Open Data,” Proceedings of International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC), pp.722–735, Nov. 2007.
[8] F.M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A Core of Semantic Knowledge,” Proceedings of International Conference on World Wide Web (WWW), pp.697–706, May 2007.
[9] F.M. Suchanek, G. Kasneci, and G. Weikum, “YAGO: A Large Ontology from Wikipedia and WordNet,” Journal of Web Semantics, vol.6, no.3, pp.203–217, Sept. 2008.
[10] 隅田飛鳥，吉永直樹，鳥澤健太郎，萬成賢太郎，“Wikipedia からの大規模な上位下位関係の獲得，” 言語処理学会第 14 回年次大会，pp.769–772，Mar. 2008.
[11] 桜井慎弥，手島拓也，石川雅之，森田武史，和泉憲明，山口高平，“汎用オントロジー構築における日本語 Wikipedia の適用可能性，” 人工知能学会第 18 回セマンティック Web とオントロジー研究会 SIG-SWO-A801-06，July 2008.
[12] 河原大輔，黒橋禎夫，“格フレーム辞書の漸次的自動構築，” 自然言語処理，vol.12，no.2，pp.109–131，Mar. 2005.
[13] D. Kawahara, and S. Kurohashi, “Case Frame Compilation from the Web using High-Performance Computing,” Proceedings of International Conference on Language Resources and Evaluation (LREC), May 2006.
[14] 数原良彦，戸田浩之，櫻井彰人，“ブログ記事を用いた複数話題語間の動作関係抽出手法，” 電子情報通信学会論文誌 D，vol.J91-D，no.3，pp.619–627，Mar. 2008.
[15] 大島裕明，小山聡，田中克己，“Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見，” 情報処理学会論文誌:データベース，vol.47，no.SIG19(TOD32)，pp.98–112，Dec. 2006.
[16] K. Nakayama, T. Hara, and S. Nishio, “A Thesaurus Construction Method from Large Scale Web Dictionaries,” Proceedings of IEEE International Conference on Advanced Information Networking and Applications (AINA), pp.932–939, May 2007.
[17] T. Kudo, and Y. Matsumoto, “Fast Methods for Kernel-Based Text Analysis,” Proceedings of Meeting on Association for Computational Linguistics (ACL), pp.24–31, July 2003.
[18] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.230–237, July 2004.
[19] K. Järvelin, and J. Kekäläinen, “IR Evaluation Methods for Retrieving Highly Relevant Documents,” Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp.41–48, July 2000.
[20] K. Eguchi, “Overview of the Topical Classification Task at NTCIR-4 WEB,” Working Notes of the 4th NTCIR Meeting, Supplement, vol.1, no.48–55, June 2004.