自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築

An Automated Web Ontology Construction Method based on NLP with Link Structure Mining for Wikipedia

中山 浩太郎 ♥ **原 隆浩** ♦
西尾 章治郎 ♠

Kotaro NAKAYAMA Takahiro HARA
Shojiro NISHIO

Wikipedia は知識抽出のための有用なコーパスとして，人工知能や情報検索，Web マイニングなどの研究分野で最近急速に注目を集めている．筆者らの研究グループでは，Wikipedia から高精度な大規模連想シソーラスを構築できることを証明してきたが，is-a 関係などのような，より明確な意味関係の抽出が技術的課題であった．本研究では，リンク構造解析による重要文抽出と，自然言語処理を利用した解析手法を提案し，意味関係を抽出することで，Wikipedia から機械可読な概念辞書を自動的に構築することを目指す．

The fact that Wikipedia is an invaluable corpus for knowledge extraction has been confirmed in various research areas such as AI, IR and Web Mining. In our previous researches, we have proved that we can extract a huge scale and accurate association thesaurus from Wikipedia. However, to construct a Web ontology from Wikipedia, extracting explicit relation types is a remaining technical issue. In this paper, we propose a method to construct a Web ontology from Wikipedia based on parsing and link structure analysis.

1. はじめに

Wikipedia は，Wiki[1] をベースにした大規模 Web 百科事典である．Wiki をベースにしているため，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り，幅広い分野の記事（概念）を網羅している．現在では，一般的な概念だけでなく，文化，歴史，数学，科学，社会，テクノロジなどの幅広い分野をカバーし，普遍的な概念から新しい概念に至るまで，非常に膨大なコンテンツが網羅されている．その記事数は既に 200 万（2007 年 12 月英語のみカウント）を超えており，世界最大の百科事典である Britannica の記事数が，全 60 巻で約 65,000 記事であることと比較した場合，実に 30 倍近い数の記事が網羅されていることになる．

Wikipedia は，この幅広いトピックの網羅性以外にも興味深い特徴をいくつか持つ．密なリンク構造はその最たる例である．Wikipedia の記事をいくつか閲覧すれば，記事同士が非常に多くのリンクで相互に参照されていることがわかる．また，URL により語彙の意味が一意に特定されている点や，リンクテキストの質の高さなども Wikipedia の特徴である．これら Wikipedia の持つ特徴は，知識抽出のコーパスとして極めて有利に働くことが各種の研究によりここ数年で急速に解明されてきた．特に，概念同士の関係度を数値化した連想シソーラスの構築に有効であることが証明されている [2, 3, 4, 5]．連想シソーラスは，概念間の関係度を数値化した辞書であり，情報検索や自然言語処理，情報フィルタリングなど幅広い分野で必要とされている．筆者らは，先行研究で Wikipedia のページ間リンクの構造を解析することで，大規模で精度の高い連想シソーラス辞書を構築し，その結果を公開してきた [2, 3]．しかし，連想シソーラス辞書は，単に概念間の関係度を数値化しているだけであるため，意味関係の抽出が技術課題であった．本研究では，リンク構造解析による重要文抽出と，自然言語処理を利用した解析手法を提案し，意味関係を抽出することで，Wikipedia から機械可読な概念辞書を自動的に構築することを目指す．

本研究には二つの技術的課題が存在する．一つ目の課題は自然言語解析技術の最適化である．明確な意味関係を抽出するためには，リンク構造解析だけではなく，自然言語解析が必要不可欠であるが，Wikipedia の記事は Wiki の特殊な構文に従って記述されるため，通常の自然言語解析を適用したときに解析精度が著しく低下するという問題がある．そのため，自然言語解析が可能な形に文章を整形，タグ付けする必要がある．二つ目の課題は，スケーラビリティである．Wikipedia の中には非常に膨大な量の文章が存在するため，すべてのセンテンスを自然言語解析することは効率的ではない．本研究では，これらの問題を考慮した解析手法を検討し，Wikipedia から大規模 Web オントロジを構築するための一手法を提案する．特に，二つの重要文抽出アルゴリズムによる精度の向上を図る．

本論文の以下では，第 2 章で関連研究について述べ，本研究のスタンスを明確にする．第 3 章では，Web オントロジの構築手法について述べる．第 4 章では Web オントロジの構築に関する実験について議論する．最後に第 5 章でまとめと今後の展開について述べる．

♥ 正会員 東京大学 知の構造化センター nakayama@cks.u-tokyo.ac.jp
♦ 正会員 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 hara@ist.osaka-u.ac.jp
♠ 正会員 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 nishio@ist.osaka-u.ac.jp

2. 関連研究

2.1 Wikipedia からの単語間関係性の抽出

Wikipedia マイニングの中でも，現在最も研究が盛んに行われているのが，概念間の関係度（Relatedness）解析である．この分野の研究としては，Strube らの研究である WikiRelate[4]，Milne らの研究 [6]，Gabrilovich らの研究 [5]，そして筆者らの研究である「Wikipedia シソーラス」[7, 8] などが挙げられる．

WikiRelate[4] や Milne らの研究 [6] は，記事が属するカテゴリ情報を利用して概念間の関係度を算出する．WikiRelate では，二つの概念が与えられた時に，カテゴリの中でその二つの概念がどれほど近いかを基準に関連度（Relatedness）を算出するという手法である．しかし，Gabrilovich らの研究 [5] で示されているとおり，カテゴリリンクを利用した手法は精度の点で問題がある．そこで，筆者らの従来研究でスケーラビリティ・精度ともに高い手法である $pfibf$[3] を提案し，その有効性を確認した．

2.2 Wikipedia 上でのオントロジ構築

Semantic Wikipedia[9] では，Wikipedia の拡張アーキテクチャとして，リンクに意味情報を付与する仕組みを提案している．例えば，通常の Wikipedia のマークアップ言語では，「'London' is the capital city of [[England]]」と記述することで，概念 London と England が関連する語であるということは記述できる．しかし，Semantic Wikipedia では，「'London' is the capital city of [[capitalof::England]]」と記述することで，London は England に対して首都である（capitalof）の関係を持つということを機械可読なメタデータとして付与する．これにより，大規模な概念体系を構築し，意味中心の WWW を実現するという計画である．Semantic Wikipedia は，大規模な Web オントロジを構築するという目的から見たときに有力な手法である可能性を持っているが，このようにユーザに新たな労力を強いるアプローチが，コミュニティに受け入れられるかという問題には，未だ疑問が残っており，小規模な概念体系しか構築できていないのが現状である．そのため，現在の Wikipedia に蓄積されている膨大なコンテンツを有効に活用し，自動的に Web オントロジを構築する手法が求められている．

3. リンク構造解析と自然言語解析による Web オントロジの構築

本章では，Wikipedia からの Web オントロジの構築に関する一手法を提案する．図 1 に本手法の概要を示す．本手法は，三つのプロセスから構成される．一つ目のプロセスでは，与えられた記事から重要文を抽出し，解析対象の文章数を制限する．二つ目のプロセスでは，Wikipedia に最適化した自然言語処理を施すことで，文章を整形し，部分的にタグを付与する（部分タギング）．三つ目のプロセスでは，抽出された重要文に対して構文解析を適用することで意味関係を抽出する．以下に各プロセスについて詳述する．

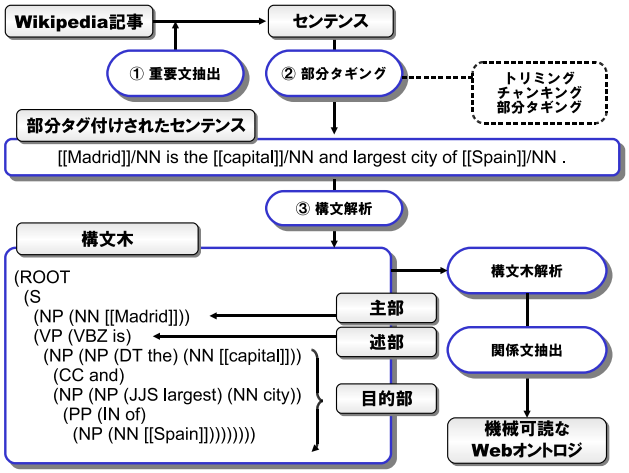

図 1 提案手法の概要

Fig. 1 Proposed Method Overview

表 1 Wikipedia のリード部分に関する統計

# of concept pages (exc. redirect and category pages)	1,580,397
# of pages having more than 100 backward links: P_a	65,391
# of pages (in P_a) begin with is-a definition sentence: P_b	56,438
# of pages (in P_a) that the 1st sentence has links: P_c	62,642
# of $P_b \cap P_c$	56,411

3.1 重要文の抽出

重要文の抽出プロセスでは，Wikipedia の記事から重要文を抽出するために LSP（Lead Sentence Parsing）法と ISP（Important Sentence Parsing）法の二つの手法を提案する．以下，二つの手法について詳述する．

3.1.1 LSP（Lead Sentence Parsing）法

LSP 法は，記事のリード部分（冒頭文）を重要文と見做して解析する手法である．これは，Wikipedia の各記事において，リード部分が多くの場合に他の概念との明確な意味関係を定義した文であることを利用した手法である．特に，Wikipedia におけるリード部分は，他の概念に対する is-a 関係が豊富に定義されていることがこれまでの調査によって判明している．リード部分に関する統計情報（2006 年 9 月のデータ）を表 1 に示す．

Wikipedia 全体では，約 158 万のページ（リダイレクトページとカテゴリページを除く）が存在するが，情報の信頼性が低い「ノイズページ」を除くためにバックワードリンク数が 100 以下のページを削除したところ，約 6 万 5 千ページを抽出できた．バックワードリンク数に応じてノイズページを除外する方

法は，Gabrilovich らの研究 [5] でその有効性が証明されている．次に，リード部分が「is-a」関係（is/are/was/were）を定義した文であるかを解析したところ，実に 86.3% (P_b/P_a) ものページが「is-a」関係を定義したページであることが判明した．さらに，95.7 % (P_c/P_a) のページは，他のページに対するリンクがリード部分に存在していた．そして，85.5 % ($(P_b \cap P_c)/P_a$) のページは，リード部分に「is-a」関係と他のページに対するリンクを保持していることが判明した．この統計情報は，Wikipedia の各ページにおいて，リード部分は，他の概念に対しての「is-a」関係を抽出するために有用な情報を含んでいる可能性が高いことを示している．

3.1.2 ISP（Important Sentence Parsing）法

ISP 法は，記事の中から重要な文章を抽出して解析する手法である．ここで，重要な文章とは，その記事の中で重要なリンクや単語を含む文章のことである．重要なリンクや単語を抽出する方法には，リンクの共起性解析や TF-IDF などの手法が利用可能であるが，本研究では，筆者らが提案する精度の高い連想シソーラスの構築手法 $pfibf$[8] を利用する．$pfibf$ は，精度の高さとスケーラビリティを兼ね備えた連想シソーラスの構築手法であり，特定の記事の中に含まれる重要なリンクを抽出することが可能である．以下に本手法の詳細を説明する．

$pfibf$ は，リンク構造解析手法であり，グラフ $G = \{V, E\}$（V はページの集合，E はリンクの集合）内において n ホップ以内のノード同士の関係性を数値化することを目的としている．ここで，Wikipedia では一つの記事（ページ）が一つの概念に対応するため，二つの記事間の関係性を抽出することは，二つの概念間の関係性を抽出することと同義である．二つの記事間（v_i, v_j）の関係の強さを計測する問題を考えた場合，関係の強さは以下の二つの要素に依存すると考えられる．

- 記事 v_i から記事 v_j へのパスの多さ
- 記事 v_i から記事 v_j への最短距離

つまり，記事 v_i から記事 v_j へのパスが多ければ多いほど（共通のリンク先や共通の参照元が多いほど），記事間の関係性は強く，またそのパスの長さが短ければ短いほど強く関係すると考えられる．v_i から v_j への n ホップ先の全経路 $T = \{t_1, t_2, ..., t_n\}$ が与えられたとき，記事 v_i から記事 v_j の関係性 $pfibf$（Path Frequency Inversed Backward link Frequency）を以下の式により表現する．

$$pfibf(v_i, v_j) = \sum_{k=1}^{n} \frac{1}{d(|t_k|)} \cdot \log \frac{N}{bf(v_j)}. \quad (1)$$

d は経路 t_k の経路長に応じて増加する関数であり，単調増加関数を利用する．N は全記事数，$bf(v_j)$ は記事 v_j が持つ他の記事からのリンク数とする．つまり，$pfibf$ は多くのリンク先を共有するが，他の記事とはリンク先を共有しない記事により高い値を示す．また，同じ距離（例えば距離 1，直接リンク関係にある）の記事であっても，より多くリンク先を共有する記事に対して高い値を示す．$pfibf$ で得られた連想シソーラスの例を表 2 に示す．

表 2 $pfibf$ によって抽出された連想関係の例

Query	Extracted association terms		
Sports	Basketball	Baseball	Volleyball
Microsoft	MS Windows	OS	MS Office
Apple Inc.	Macintosh	Mac OS X	iPod
iPod	Apple Inc.	iPod mini	iTunes
Book	Library	Diamond Sutra	Printing
Google	Search engine	PageRank	Google search
Horse	Rodeo	Cowboy	Horse-racing
Film	Actor	Television	United States
DNA	RNA	Protein	Genetics
Canada	Ontario	Quebec	Toronto

ISP 法では，$pfibf$ で得られた概念間の連想関係を利用して，ページ（概念）の中で重要なリンクと単語を含むセンテンスを抽出し，解析対象とした．

3.2 部分タギング

部分タギングプロセスでは，Wiki の特殊な文法で記述されたテキストを自然言語処理可能な状態に整形する．本ステップは，大きく分けて，1）トリミング，2）チャンキング，3）タギングの手順で処理を行う．

トリミングでは，HTML タグなどの不要な情報を削除する．さらに，Wiki の特殊タグも削除する．特殊タグには，テーブルタグなどが含まれる．テーブルタグは，多くの場合はセンテンスが含まれず，提案手法が適用できないため，削除の対象とする．しかし，Wiki の特殊タグの中でも，リンクに関するタグは削除しない．これは，リンクのタグが本研究で重要な情報となるためである．次に，チャンキングでは，構文解析のためにテキストをセンテンス単位に分割する．通常，このような処理には，チャンキングのツール（OpenNLP[10] など）を利用するのが一般的であるが，Wiki の特殊なタグが含まれる状態では，正常にテキストを分割することができない．そのため，本研究では，改行コードやピリオド（.），空白文字などの情報を利用して，単純なルールでセンテンスを分割する．最後に，(ダブル）クォーテーションされている単語や，リンク部分は一つの名詞としてまとめると共に，自然言語処理を助けるために POS（Part Of Speech）タグを付与する．POS タグは，自然言語処理研究において語に付与するラベルであり，代表的な POS タグには表 3 に示すようなものがある．

3.3 構文解析

構文解析のステップでは，上述のステップで部分的にタギングされた文章に対して構文解析を行った．構文解析には，構文解析手法として良く利用される確率的構文解析法を適用した．本研究ではこの手法を再現するために，Stanford NLP Parser[11] を利用した．Stanford NLP Parser は，与えられたセンテンスの

表 3　主な POS タグ

Tag	Description
S	Sentence
NN	singular or mass noun
NNS	Plural noun
NNP	Singular proper noun
NNPS	plural proper noun
NP	Noun phrase
VB	Base form verb
VBD	Past tense
VBZ	3rd person singular
VBP	Non 3rd person singular present
VP	Verb phrase
JJ	Adjective
CC	Conjunction, coordinating
IN	Conjunction, subordinating

構文を解析し，POS タグが付与された構文木を生成する．例えば，以下のようなセンテンスからは，

```
Lutz_D._Schmadel is [[Germany|German]] [[astronomer]]
```

以下のような構文木が生成される．

```
(S (NP (NN Lutz_D._Schmadel)
    (VP (VBZ is)
     (NP (NN [[Germany|German]]) (NN [[astronomer]])
))))
```

提案手法では，生成された構文木から以下の手順に従って意味情報を抽出する．

1. 「(NP ...) (VP (VBZ/VBD/VBP ...) (NP ...))」パターンを抽出する．
2. もし，NP が JJ と NN/NNS からのみ構成されている場合，NP を最後の NN/NNS で置き換える．
3. もし，NP が CC を含んでいる場合，CC の場所で NP を二分割する．そして，ステップ 2 を再度実行する．
4. 最後に，最初の NP を主語，二つ目の NP を目的語，VP を述語とする意味的三つ組みを抽出する．

例えば，上述の「Lutz_D._Schmadel」に関するセンテンスでは，二つ目の NP は二つの NN から構成され，そのどちらもリンクを保有する．一つ目の NN はページ「Germany」へのリンクであり，もう一つはページ「astronomer」へのリンクである．この場合，最後の NN である「astronomer」が NP の語幹であるため，NP 全体を「astronomer」で置き換える．そして，最後に「Lutz_D._Schmadel」「is」「astronomer」を意味的三つ組みとして抽出する．

表 4　評価結果

Method	Literal	Extracted Relations	Correct Relation	Precision
ASP	Includes	458	285	62.22 %
	Excludes	162	133	82.09 %
LSP	Includes	101	91	90.09 %
	Excludes	54	52	96.30 %
ISP	Includes	67	54	80.59 %
	Excludes	59	51	86.44 %
$LSP \cup ISP$	Includes	153	130	84.96 %
	Excludes	99	88	88.88 %

本研究では，概念の多義性を考慮した意味関係の抽出を目指しているため，二つ目の NP にリンクが含まれる場合を主な対象としている．これは，前述のとおり Wikipedia では一つのページが一つの概念に対応し，リンクを利用することで多義性を解消した意味関係抽出ができるためである．しかし，その一方で，NP が文字列（リテラル）のみで構成される場合の情報もアプリケーションによっては有用であると考えられる．例えば，情報検索などのアプリケーションでは，精度よりも網羅性の方が重要視される場合もあるためである．そのため，本研究では NP が文字列で構成されている時にも意味関係を抽出し，リンクで構成されているときとは別に精度を計測した．

4. 評価実験

4.1 実験概要

本研究では，まずはノイズ記事をフィルタするために，バックワードリンク数が 100 以上のページをすべて収集し，上述のアルゴリズムに基づいて意味解析を行い，Web オントロジの構築を試みた．その後，ランダムに 110 の記事を抽出し，1,016 のセンテンスから抽出した意味関係のセットを評価した．ここで，二つの構文解析戦略（LSP 法と ISP 法）のほかに，比較対象として全文を解析する手法（ASP: All Setence Parsing 法）も適用した．

4.2 実験結果と考察

解析結果を表 4 に示す．

ここで，まず注目すべき点は，LSP 法の精度の高さである．LSP 法は，リンクを利用した明確な意味関係抽出とリテラルを利用した意味関係抽出の両方で高い精度を実現している．これは，意味関係を抽出する上で有用な情報がリード部分に多く含まれているという本研究の仮定が正しかったことを示している．Wikipedia では通常記事の上部に書かれている内容は，下部に書かれている内容より信頼できることが多いという特性を持っている．これは，記事の上部に書かれている内容は，多くのユーザによって閲覧され，内容に間違いがあったときに即座に修正されるためである．そのため，リード部分では，多くの有用な知識が抽出できたと考えられる．

表 5　LSP 法によって抽出された明確な意味関係の例

Subject	Predicate	Object
Apple	is	Fruit
Bird	is	Homeothermic
Bird	is	Biped
Cat	is	Mammal
Computer	is	Machine
Isola_d'Asti	is	Comune
Jimmy_Snuka	is	Professional_wrestler
Karwasra	is	Gotra
Mineral_County ,_Colorado	is	County
Nava_de_Francia	is	municipality
Sharon_Stone	is	Model
Sharon_Stone	is	Film_producer

表 6　ISP 法によって抽出された明確な意味関係の例

Subject	Predicate	Object
Isola_d'Asti	borders	Asti
Isola_d'Asti	borders	Costigliole_d'Asti
Isola_d'Asti	borders	Mongardino
Mauritania	is	Nouakchott
San_Francisco_Peninsula	cross	San_Francisco_Bay
Sharon_Stone	hosted	Nobel_Peace_Prize_ Concert
Karl_Ziegler	was	Chemist
Lutz_D._Schmadel	is	Astronomer
People's_Party_for _Freedom_and_Democracy	is	political_party
San_Francisco_Peninsula	separates	San_Francisco_Bay
Tom_Hicks	is	Businessman

次に注目すべき点は，ISP 法も LSP 法と同様に，ASP 法に比べて精度の面で良い成果を出していることである．特に，リテラルを利用した意味関係抽出では，ISP 法が ASP 法に比べて著しく高い精度を実現している．さらに，ISP 法では構文解析を行う前段階で重要な文を選定するため，ASP 法に比べて計算時間を大きく短縮できている点も重要である．また，LSP 法と ISP 法を組み合わせることで，概念の網羅性と精度を同時に実現することができている．

表 5 に LSP 法で抽出した明確な意味関係の例を示す．ここで，明確な意味関係とは，リンクによって多義性が解消された状態の意味関係を示す．表からは,「is-a」の意味関係が精度良く抽出できていることがわかる．これは，前述のとおりリード部分が「is-a」関係を豊富に保持しているためである．

表 6 に，ISP 法によって抽出された意味関係の例を示す．ISP 法は，その記事にとって重要な概念を抽出する手法であるという特性から，様々な種類の関係（borders や hosted）が抽出できている．しかし，機械理解可能な Web オントロジを実現するためには，これらの関係を抽出するだけでは不十分であり,「関係同士の関係」をさらに定義する必要がある．例えば,「was は is の過去形である」といった関係同士の関係を定義することで，抽出された意味関係を機械処理可能な形式へと変換することが可能である．

最後に，表 7 に ISP と LSP によって抽出したリテラルの意味関係の例を示す．リテラルの意味関係を抽出した結果から判明したことは，目的語がリテラルとして定義されるのは，目的語が「city」や「town」など一般的な語の場合が多いということである．これは，Wikipedia においてはこのような一般語に対してはリンクを付与する場合が少ないことに起因すると考えられる．このようなリテラルに対しては，さらに意味の推定を行う必要がある．

表 7　ISP 法と LSP 法によって抽出されたリテラルの意味関係の例

Subject	Predicate	Object
Taranto	is	Coastal city
The_Isley_Brothers	is	Black music group
Toronto_Islands	is	Chain
Mauritania	is	Country
Mauritania	is	Country
Ilirska_Bistrica	is	Town
Ilirska_Bistrica	is	Municipality
Brescia	is	City

4.3　議論と今後の展開

本実験では，Wikipedia のセンテンスに対して構文解析を行ったが，技術的課題として,「照応解析」の高度化が必要であることがわかった．照応解析とは，ある一文が入力として与えられた時に，その主語部分が指し示す言葉の意味を特定する処理である．たとえば，Wikipedia の記事の中では，記事の主題が最初に出現する場所では省略せずに記載されるが，二回目以降は省略形や代名詞，同義語が利用されるのが一般的である．そのため，各センテンスの主語部分が示す単語の意味を特定することで，概念関係抽出の網羅性を高める必要がある．

照応解析には，１）頻出代名詞を主題の照応とする手法や２）記事のタイトルに含まれる語を主題の照応とする方法などが従来研究で提案されている [12] が，省略形や同義語などを考慮して解析することで精度の向上が期待できる．

5.　まとめと今後の課題

本研究では，自然言語解析とリンク構造解析を利用することで，Wikipedia から Web オントロジを自動生成する手法を提案した．また，実験の結果から，各記事のリード部分には有用な知識が蓄えられており，特に is-a 関係を高精度に抽出するという用

途に向いていることが判明した．さらに，ISP 法では is-a 関係だけでなく，複雑な関係の抽出が可能であることが判明した．本研究の成果は，既に以下の URL で公開している．

- Wikipedia Lab.
  http://wikipedia-lab.org
- Wikipedia Thesaurus
  http://wikipedia-lab.org:8080/WikipediaThesaurusV2
- Wikipedia Ontology
  http://wikipedia-lab.org:8080/WikipediaOntology

Wikipedia Ontology の Web サイトでは，LSP 法によって抽出された約 32,000 の意味関係が公開されている．今後の展開としては，より詳細にスケーラビリティや精度の評価を行い，どのようなパラメータが精度に影響を与えるかを調査する．また，自然言語解析のパターンを増やすことで，網羅性と精度の向上を図ることも重要な技術的課題である．さらに，利用頻度の高い関係同士の関係を定義することで，機械処理可能な意味関係のネットワークを構築可能することを目指す．

謝辞：本研究の一部は，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] W. C. Bo Leuf: “The Wiki Way: Collaboration and sharing on the Internet”, Addison-Wesley (2001).

[2] K. Nakayama, T. Hara and S. Nishio: “A thesaurus construction method from large scale web dictionaries.”, Proc. of IEEE International Conference on Advanced Information Networking and Applications (AINA 2007), pp. 932–939 (2007).

[3] K. Nakayama, T. Hara and S. Nishio: “Wikipedia mining for an association web thesaurus construction.”, Proc. of IEEE International Conference on Web Information Systems Engineering (WISE 2007) (2007).

[4] M. Strube and S. Ponzetto: “WikiRelate! Computing semantic relatedness using Wikipedia”, Proc. of National Conference on Artificial Intelligence (AAAI-06), Boston, Mass., pp. 1419–1424 (2006).

[5] E. Gabrilovich and S. Markovitch: “Computing semantic relatedness using wikipedia-based explicit semantic analysis.”, Proc. of International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 1606–1611 (2007).

[6] D. Milne, O. Medelyan and I. H. Witten: “Mining domain-specific thesauri from wikipedia: A case study”, Proc. of ACM International Conference on Web Intelligence (WI’06), pp. 442–448 (2006).

[7] 中山, 原, 西尾：“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, **47**, 10, pp. 2917–2928 (2006).

[8] 中山, 原, 西尾：“Web 事典からのシソーラス辞書構築手法”, 情報処理学会論文誌: データベース, **48**, SIG19(TOD 34), pp. 27–37 (2007).

[9] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller and R. Studer: “Semantic wikipedia”, Proc. of International Conference on World Wide Web (WWW 2006), pp. 585–594 (2006).

[10] OpenNLP forum: “Opennlp”, http://opennlp.sourceforge.net/ (2006).

[11] D. Klein and C. D. Manning: “Accurate unlexicalized parsing”, Proc. of Meeting of the Association for Computational Linguistics (ACL 2003), pp. 423–430 (2003).

[12] D. P. T. Nguyen, Y. Matsuo and M. Ishizuka: “Relation extraction from wikipedia using subtree mining”, Proc. of National Conference on Artificial Intelligence (AAAI-07), pp. 1414–1420 (2007).

中山 浩太郎 Kotaro NAKAYAMA

2003 年関西大学大学院総合情報学研究科修士課程修了．この間 (株) 関西総合情報研究所代表取締役，同志社女子大学非常勤講師に就任．2007 年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博士後期課程修了後，同大学院情報科学研究科の特任研究員を経て，2008 年東京大学知の構造化センター特任助教に就任．人工知能および WWW からの知識獲得に関する研究に興味を持つ．ACM，IEEE，情報処理学会，電子情報通信学会，日本データベース学会の各正会員．

原 隆浩 Takahiro HARA

1997 年大阪大学大学院工学研究科博士前期課程修了．同年，博士後期課程中退後，同大学大学院工学研究科情報システム工学専攻助手，同大学大学院情報科学研究科マルチメディア工学専攻助手を経て，2004 年より同大学大学院情報科学研究科マルチメディア工学専攻准教授となり，現在に至る．工学博士．データベースシステム，モバイルコンピューティングなどの研究に従事．IEEE，ACM，電子情報通信学会，情報処理学会，日本データベース学会の各会員．

西尾 章治郎 Shojiro NISHIO

1980 年京都大学大学院工学研究科博士後期課程修了．工学博士．京都大学工学部助手，大阪大学基礎工学部および情報処理教育センター助教授，大阪大学大学院工学研究科教授を経て，2002 年より同大学大学院情報科学研究科教授となり，現在に至る．2000 年より大阪大学サイバーメディアセンター長，2003 年より大阪大学大学院情報科学研究科長を併任．2007 年より同大学副学長理事を併任．データベース，マルチメディアシステムの研究に従事．現在，Data & Knowledge Engineering 等の論文誌編集委員，本学会理事，電子情報通信学会，情報処理学会の各フェローを含め，ACM，IEEE など 8 学会の会員．