情報処理学会論文誌　データベース　Vol. 2　No. 4　49–60 (Dec. 2009)

Wikipediaマイニング：Wikipedia研究のサーベイ

中山　浩太郎[†1]　伊藤　雅弘[†2]　Erdmann Maike[†2]
白川　真澄[†2]　道下　智之[†2]
原　隆浩[†2]　西尾　章治郎[†2]

Wikipediaは，インターネットを通じて誰でも編集可能なオンライン百科事典であり，ここ数年で爆発的に成長したソーシャルメディアの一種である．特に，自然言語，人工知能，データベースの研究分野で活発に研究が進められており，連想関係抽出や，対訳辞書構築，オントロジ構築など，数多くのWikipediaを対象とした研究が行われてきた．また，最近では多様なアプリケーションへWikipediaマイニングの成果を適用する事例が報告されており，その有用性が示されてきた．しかし，多量の研究発表が行われる一方で，全体像を把握することが困難になりつつあるのも事実である．本サーベイ論文では，これら最新のWikipedia研究を紹介しつつ，概観することで研究の目的面・技術面から分類し，Wikipedia研究の動向を探る．

Wikipedia Mining: A Survey on Wikipedia Researches

Kotaro Nakayama,[†1] Masahiro Ito,[†2]
Maike Erdmann,[†2] Masumi Shirakawa,[†2]
Tomoyuki Michishita,[†2] Takahiro Hara[†2]
and Shojiro Nishio[†2]

Wikipedia, an Wiki based online encycropedia, has become an emergent social media because of the significant effeciency for sharing huge amount of human knowledge via Web browsers. Especially, in NLP, AI and DB research areas, a considerable number of researches have been conducted in past several years. Relatedness measurement, bilingual dictionary extraction and ontology construction are ones of main Wikipedia Mining research areas. Furthermore, researches on application based on structured data extracted by Wikipedia Mining are becoming one of the essentials of Wikipedia research areas. In this survey paper, we introduce the new research papers and summarize the researches from both technical aspect and directional aspect.

1. はじめに

Wikipediaは，インターネットを通じて誰でも編集可能なオンライン百科事典であり，ここ数年で爆発的に成長したソーシャルメディアの一種である．Wikipediaは，Webブラウザを通じて自由に編集可能なことから，ユーザ同士が迅速かつ容易にコンテンツを編集するための情報基盤を提供してきた．この結果，多くのユーザが精査することによって質の高いコンテンツの量は増え，さらなるユーザを獲得するというサイクルを形成している．

インターネット上に新しい情報共有の基盤を作り上げたWikipediaは，社会現象としても興味深いが，研究者にとっては魅力的な新しい研究用のリソースへと成長してきた．これを証明するように，ここ数年でWebの研究分野をはじめ，人工知能や自然言語処理，情報検索など幅広い研究分野で研究の基盤リソースとして利用され，その有用性が示されてきた．特に，統計的手法に基づく自然言語処理の研究領域では，ある程度まとまった量の高品質なテキスト情報がコーパスとして必要であったが，Wikipediaはこの要件を満たし，GFDL（GNU Free Documentation License）に基づくコピーレフトなライセンス形体で利用しやすいという点から，標準的な言語リソースの1つになりつつある．

Wikipediaに関する研究領域は多岐にわたり，概念どうしの関係性を数値化する連想関係抽出，より詳しい関係の種類を抽出するオントロジ構築，さらには語の曖昧性解消アルゴリズムの基盤情報としての利用や情報検索への応用などに関して研究が進められてきた．ここで，Wikipediaマイニングとは，Wikipediaを解析することで有用な情報抽出を行う研究を総称するものとする．

筆者らは，「人工知能研究の新しいフロンティア：Wikipedia」[30]において，Wikipediaが研究のためのリソースとして有用であることを示したが，その後，Wikipediaマイニングに関する研究が急速に増えてきた．また，アプローチや手法，研究の方向性も，多様化が進み，全体像を把握することが困難となってきた．Wikipediaを研究対象とする論文を集めた「Wikipedia in academic studies」[⋆1]に公開されている論文だけでも，全部を読み，トレンドを把握するのは困難な状況である．そこで本稿では，これらの論文のうち，新規性・

†1 東京大学知の構造化センター
The Center for Knowledge Structuring, The University of Tokyo
†2 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
⋆1 http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies

有用性の両面で興味深いと思われるものを筆者らの視点でまとめた．本稿が今後の研究者の一助となれば幸いである．以降，Wikipedia の言語リソースとしての特徴のうち重要なものを簡単に解説した後，Wikipedia マイニング研究を技術・目的の両面から分類し，俯瞰・考察する．

2. Wikipedia の特徴

「Wikipedia」は，Wiki[12] 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る．この編集の容易さがインターネットユーザの書き込みを促進し，今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．その記事数は英語版ですでに 300 万，日本語版でも 60 万（2009 年 8 月）に達しており，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 46 倍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研究の対象として見たときに，URL による概念識別，270 を超える言語のサポート，半構造化データなど興味深い特徴をいくつか持つ[30]．詳しい特徴については文献 30) に委ねるが，以降，本章では Wikipedia のコーパスとしての特徴について特に重要な部分を略解する．

2.1 URL による概念識別（Disambiguation）

URL により語彙の一意性が確立されている点は，Wikipedia の大きな特徴の 1 つである．通常の自然言語処理の精度を低下させる要因の 1 つに語の曖昧性回避問題がある．通常のコーパスでは，曖昧性の高い単語に対してコンテキストを解析しながらその曖昧性を解消するのが一般的だが，曖昧性回避の精度がその後の処理に影響を与えるため，ある程度の精度で解析できたとしても最終的な解析の精度を下げてしまう要因になっていた．

一方，Wikipedia では 1 つの URL（ページ）に 1 つの概念が割り当てられており，多義性が URL によって解決されている点が特徴である．たとえば，「Apache」は強いコンテキスト依存性を持つ単語であり，先住民族を示す場合も HTTP サーバや軍用ヘリを示す場合もある．Wikipedia では，これら 3 つの概念は別のページで管理されており，それぞれ「http://en.wikipedia.org/wiki/Apache」「http://ja.wikipedia.org/wiki/Apache_HTTP_Server」「http://en.wikipedia.org/wiki/AH-64_Apache」という別々の URL が割り当てられている．

このように，概念と URL が 1 対 1 で対応していることは，概念の関係を解析する際に多義性やコンテキスト依存性の影響を受けずに解析できることを示している．

2.2 カテゴリリンク

カテゴリリンクは，ある記事（概念）がどのようなカテゴリに属するかを指定するためのリンクである．すべてのカテゴリには専用のページ（カテゴリページ）が用意されており，カテゴリページはさらに別のカテゴリページに属することが可能である．このカテゴリ構造は，一種のタクソノミー（分類辞書）としての役割を有しており，カテゴリを絞り込みながら記事を検索するような機能を実現するために利用されている．Wikipedia が提供しているカテゴリ検索システム「CategoryTree」[⋆1]（以降カテゴリツリー）では，カテゴリを検索することや，カテゴリの階層構造をブラウジングすることが可能である．Wikipedia の英語版（2009 年 8 月）を調査したところ，約 3,570 万のカテゴリリンクが存在していた．

しかし，一見ツリー構造に見える Wikipedia のカテゴリ構造は，実は複雑なネットワーク構造をしている．1 つのカテゴリページは複数のカテゴリページに属することが可能であり，一部にはループなども存在する．そのため，Wikipedia のカテゴリを利用する際には，このことを考慮して解析をしなければならない．

2.3 半構造化データ

Wikipedia にはカテゴリリンクやリダイレクトリンクなど，いくつかの半構造化されたデータが存在する．これらの情報は，比較的容易に解析可能なデータ形式をしているうえに，情報量が豊富であるため，分類的関係や同義語関係の抽出などの研究によく利用されている．さらに，「インフォボックス」も解析が容易な半構造化データとして，Wikipedia 研究では頻繁に利用される．インフォボックスは，各記事において，属性情報を記述するためのテンプレートであり，人に関する記事であれば「生年月日」「血液型」「身長」といった属性情報が記載される．また，都市や国に関する記事であれば，「首都」や「隣接する国」といった属性情報が記述される．インフォボックスは，Wiki のテンプレート機能を利用して記述されており，構文が明確に定義されているため，自然言語処理などを利用しなくても情報を抽出できる．そのため，事象とその属性情報を抽出し，トリプルを抽出するために利用されることが多い．

2.4 270 を超える言語サポート

Wikipedia は 2009 年 8 月の段階ですでに 270 を超える言語をサポートしており，各言語

⋆1 http://en.wikipedia.org/wiki/Special:CategoryTree

においても広い範囲の記事が網羅されている．Wikipedia では同じ概念を記述した異なる言語のページは，言語間リンク（Interlanguage Link）と呼ばれる対訳関係を示す特殊なリンクによって結ばれる．日本語と英語の記事の間には約 30 万件の言語間リンク（2009 年 8 月）が存在しており，和英電子辞書の EDICT が 12 万件であることと比較したときに十分な数の対訳関係が存在することが分かる．一般的に，（電子）辞書は限られた数の専門家によって更新されるため，最新概念に対する網羅性が弱いという問題があるが，Wikipedia は最新の概念や専門用語に対しての網羅性が高いことから従来の辞書の弱点を補完できるのではないかという期待もある．

3. Wikipedia 研究の分類

Wikipedia に関する研究は，大きく 2 つの分野に分類できる．1 つは，Wikipedia を社会現象として解明しようとする研究である．これは，たとえば Wikipedia に参加する人の目的や行動内容を調査し，社会現象としての Wikipedia を説明しようといったような研究が該当する．

もう一方の研究分野は，Wikipedia を言語リソースとして利用・解析する研究である．たとえば，概念間の関係性などの有用な情報を抽出し，さらに情報検索などのアプリケーションに適用するような研究がこの分野に分類される．本稿では，後者の研究分野を Wikipedia マイニング研究として対象とする．

Wikipedia マイニングに関する研究は，様々な目的・アプローチが混在し，一様に分類することは難しいが，本稿では技術的な側面から，以下の 4 つのカテゴリに分類した．

- 関連度計算（Relatedness Measurements）
- 関係抽出（Relation Extraction）
- 対訳辞書構築（Bilingual Dictionary Construction）
- アプリケーション

以降，各研究分野について詳述する．

4. 関連度計算（Relatedness Measurements）

概念の関連度を数値化する研究は，情報検索のクエリ拡張や文書分類など利用用途が広いことから，情報検索や文書分類の分野で広く研究が進められてきた．また，Wikipedia マイニングの研究の中でも，最も歴史が長く，活発に行われている研究分野の 1 つである．関連度計算とは，任意の概念ペア間の関連度の強さを数値として算出する処理である．たとえば，「コンピュータ」と「メモリ」などは比較的高い関連性を持つ概念だが，「コンピュータ」と「ジャガイモ」などは一般的に関連度が高いとはいえない．本研究領域では，このような関連の強さを数値化することを目的としている．また，2 つの概念からその関連度を計算するだけでなく，1 つの概念から関係の深い概念集合を抽出するような研究もこの研究分野に分類される．

Wikipedia マイニングにおける関連度計算で重要なポイントは，Wikipedia では URI によって概念が一意に特定されるという点である．この結果，同じ Apple でもフルーツの「Apple」と「Computer」の関連度は低いが「Apple Inc.」と「Computer」の関連度は高いなど，語の曖昧性が除去された状態で，単語ではなく概念ごとの関連度が得られる．これは，従来の自然言語処理で困難な問題の 1 つとされていた語の曖昧性問題が解決されている状態で関係度の計算を行えることを意味している．

Wikipedia をリソースとして利用することにより，従来の言語リソースの解析では実現できなかった高い網羅性を実現できると期待されている．これは，WordNet では固有名詞や専門用語などは網羅されていないことが課題であったが，Wikipedia ではこれらの語が多く網羅されており，新しい語にも迅速に対応されるためである[25]．

また，関連度計算では関連性の強い概念ペアが抽出できるため，次のステップとしてそれらの概念が具体的にどのような関係（例：is-a，has-a）にあるか，という関係抽出の研究の基本情報として利用することが可能である．

関連度計算の手法としては，従来の研究で WordNet や Web 上のドキュメントなどに適用されてきた概念階層を用いた手法やテキストの一致度を用いた手法を，改めて Wikipedia に適用したものが多かったが，最近では，Wikipedia 独自の構造（アンカーテキストなど）を用いた手法が研究されてきている．本章では，各研究を分類しながら解説する．

4.1 カテゴリツリーの解析手法

Wikipedia マイニングにおいて，最も一般的な関連度計算のアプローチの 1 つがカテゴリツリーを利用する方法である．前述のとおり，カテゴリツリーは概念を分類するための階層的構造であるが，本アプローチはカテゴリツリーにおいて概念間のパスの長さが短いほど関連度が高くなる，という考えに基づいている．

代表的な研究として，Strube らの WikiRelate![25] があげられる．Strube らは，WordNet に用いられてきた関連度算出の手法が Wikipedia に適用できることを証明し，複数の指標を統合することで精度が向上することを示した．WikiRelate!では，カテゴリツリーの解析手法をさらに 3 つの手法に分類している．1）カテゴリ構造におけるパスの長さに基づく手

法，2）カテゴリ構造における情報の共有度（直近の共通の祖先が持つ子概念が少ないほど関連度が高い）に基づく手法，3）記事の内容（出現単語のヒストグラムなど）のオーバラップの程度に基づく手法がある．WordSimilarity 353 テストコレクションなどを利用した実験結果では，Wikipedia を利用した手法が WordNet に匹敵する評価を残し，Wikipedia が知識抽出のためのコーパスとして有用であることが示された．

続けて Strube らは文献 21) で，文献 25) をもとに 2 つの単語の関連度を計算する API を構築している．これは，2 つの単語を入力すると，文献 25) の手法に基づき，その単語の関連度が出力され，カテゴリ構造におけるパスも視覚的に表示されるシステムである．現在，英語，ドイツ語，フランス語，イタリア語をサポートしており，オンラインで利用可能である．

4.2　テキスト内容の比較による手法

2 つ目の関連度計算手法は，テキスト内容を比較し，その類似度を利用する手法である．テキストを用いた手法は，概念に関する説明文（記事）が充実している場合に有効な手法であり，一般にテキストに出現する単語が重複していればいるほど関連度が高くなるという戦略に基づく手法である．しかし，言語によっては高度な自然言語処理が必要であり，特に日本語では形態素解析などが精度に大きく影響するといった面も持つ．

Gabrilovich らの研究[8] では，単語やテキストの意味を表現するための手法として，ESA（Explicit Semantic Analysis）を提案している．ESA では，特定の単語（文字列）またはテキストの意味を，Wikipedia の概念を基底とする高次元ベクトルで表す．単語の意味を表すベクトルは，各概念を tfidf で重み付けられた単語のベクトルで表した後，これらのベクトルの逆索引を作成することで得られる．単語の関連度は，ベクトル間のコサイン相関によって求められる．テキストの意味を表すベクトルも，出現するすべての単語のベクトルを合成することで求められ，文脈を考慮した語義曖昧性解消が可能となる．

また，前節で述べた Strube らの研究[25] では，カテゴリツリーを用いた手法だけでなく，テキストの重複度に基づく指標も関連度計算に有効であると報告しているが，これもテキスト内容の比較による関連度抽出の手法の一種である．

4.3　ページ間リンクの解析手法

Wikipedia は Wiki をベースにしており，記事の中に他の概念（を意味する単語）が出現するとその概念に対してリンクが張られるため，全体としてみると，概念をノード，ハイパーリンクをリンクとした一種のネットワークと見なすことができる．通常の Web サイトと異なり，ノード（ページ）は概念を表し，リンクは意味的な関係を表すうえに，Wikipedia 内部で概念どうしが密なリンク構造を形成している（内部リンクが多い）ため，リンク構造を解析することで概念間の関係性を抽出することが可能である．

この特徴を生かし，リンクの構造を解析して関連度計算を行うのがページ間リンクの解析手法である．この分類において，主な手法としては，ネットワーク構造における 2 つの概念のリンク数やホップ数に基づく手法，記事内における概念の共起性に基づく手法などがある．

Nakayama らの研究[17] では，pfibf という手法を提案し，大規模なシソーラスを構築している．この手法では，概念をノードとしたネットワーク構造において，概念間のパスの数が多いほど，またそれらのパスが短いほど，それらの概念が強く関連していると見なす．それに加えて，ある概念が他の概念からリンクを多く張られている（バックワードリンク数が多い）ほど，その概念が一般的な概念であるとして，関連を弱くする．さらに Nakayama らの研究[17] では pfibf に加えて，フォワードリンクによる解析とバックワードリンクによる解析の比重を可変にする手法も提案している．これは，一般的な概念と専門的な概念では，記事の信頼度やフォワードリンク・バックワードリンクの重要性が異なるためである．この手法では，従来手法である TF-IDF と比べて計算量は増加するが，精度は向上している．

Ito らの研究[10] では，記事内に出現するリンクの共起性から概念の関連度を計算している．pfibf では数ホップ先までリンクをたどる必要があるのに対し，この手法ではリンクをたどる必要がないため，pfibf と同等の精度を実現しつつも計算量が大幅に改善されている．

また Ollivier らの研究[19] では，Wikipedia に特化した手法としてではなく，一般的なグラフ中の関連ノードを発見するための手法として，マルコフ連鎖に基づく Green Measure を用いている．Green Measure はもともと静電気のポテンシャルを計算するための理論である．Ollivier らはこの理論を，PageRank のように，Web ページとハイパーリンクから構成される有向グラフに適用している．評価実験では Wikipedia を対象にしており，文章ベクトルのコサイン相関や共参照，PageRank といった従来手法と比べて，全体的に精度が高いうえに記事によって精度が劣ることがなく，従来手法で抽出できなかった関係も発見できることが示された．また，本手法はグラフの構造のみを用いており，記事やカテゴリといった他の情報を解析することで，さらに精度を向上させることができる可能性がある．

4.4　問題点と課題

カテゴリツリーの解析手法は，WordNet などのカテゴリツリーに用いられてきた既存手法をそのまま適用可能であるが，Wikipedia の記事やページ間リンクの膨大な情報量と比較すると，カテゴリツリーの情報量は少なく，精度・網羅性向上の余地がある．また，テキ

スト内容の比較による手法は，記事のノイズデータや自然言語処理にともなう精度低下の問題がある．一方，ページ間リンクの解析手法は，リンク先の概念を端的に表したアンカーテキストにより，ノイズデータを削減できるため，先にあげた 2 つの手法よりも精度の良い関連度計算が期待できる．関連度計算の手法は，ページ間リンクをベースにした手法が中心になってきている．

また，関連度計算に関する研究全体としては，関係抽出やオントロジマッピングを目標とした研究の一部分としての位置づけである場合が多い．そのため，多くの論文では今後の課題として，概念間の関係（例：is-a，a-part-of）を抽出することが重要であると述べている．実際に Strube らは，関連度計算に関する論文[25] の後，関係抽出に関する論文[22] を発表している．

5. 関係抽出（Relation Extraction）

Wikipedia 研究の中で最も活発な研究分野の 1 つが関係抽出である．関連度計算の研究では，概念間の関係性の強さを連続的な数値で表現するのに対し，関係抽出に関する研究では，概念間の明示的な意味関係を抽出することを目的としている．たとえば，2 つの概念が与えられたときに，その間の関係の強さを求めるのではなく，is-a 関係なのか part-of 関係なのか，といったような関係のタイプを抽出することを目的としている．

Wikipedia から概念間の関係を抽出する研究の方向性は，Semantic Web の目標である「意味中心の Web」を実現する基盤技術として必要な大規模 Web オントロジを実現する現実的な方法として注目されている．関係抽出の研究は主に，テキストを利用する方法，カテゴリリンクを利用する方法，インフォボックスを利用する方法に分類される．以降，各分類について詳述する．

5.1 テキストを利用する方法

Wikipedia の記事には，画像やメニューなど様々な情報が含まれるが，大部分を占めるのはテキスト部分（テキストによって記述される本文）である．テキスト部分には，リンクを利用してほかの記事（概念）との関係が定義されているため，その内容を精度良く解析できれば，多量の概念関係が抽出可能であると考えられる．テキストを解析することで概念間の関係を抽出する研究としては，PORE[29]，Nguyen らの研究[18]，Culotta らの研究[5] などがあげられる．

PORE[29] は，POL（Positive Only Learning）[13] をベースにした関係抽出のアルゴリズムを提案している．POL は，Espresso[20] に代表されるようなブーストラッピング手法の一種であり，少量の教師データをシード（種）として徐々に正解集合を拡張していく手法である．PORE では，この手法を利用しカテゴリやインフォボックスなどの構造化データを抽出しやすい部分から正解集合を作成した後に，テキストに含まれる共起リンク間の関係を抽出する際の教師データとして利用する方法を提案している．その際には，テキストからの関係抽出手法として，1 つの文中に含まれる 2 つのリンク（概念）間に挟まれるテキストを述語として抽出する方法を提案している．たとえば，「In the film “Heavenly Creatures” directered by Peter Jackson. . .」という文からは，概念「Heavenly Creatures」と「Peter Jackson」の間には「directered by」の関係があるというトリプルを抽出している．

Nguyen らの研究[18] も，PORE[29] と同様に Wikipedia の記事を解析し，2 つのエントリ間に含まれるテキストを述語として抽出するアプローチである．Nguyen らの研究では，単にエントリ間のテキストを抽出するだけでなく，構文解析の結果（構文木）に含まれるパターンを発見することで精度向上を図っている．また，照応解析の方法として，頻出代名詞を利用している．Wikipedia の記事には，通常多くの代名詞や省略表現が利用されているため，文の主題が記事のメイントピックと同一かを判定する照応解析は重要なタスクの 1 つである．Nguyen らの研究では，1 つの記事内に含まれる最頻出の代名詞を抽出し，主題と同一であると見なす手法を提案している．

Culotta らの研究[5] では，関係の抽出ルールをデータマイニングによって自動的に抽出する方法を提案している．この手法では，従来は手動で構築していたような推論ルール（例：父親（母親）の兄弟（姉妹）の息子（娘）はいとこである）を，関係抽出とデータマイニング手法を統合するアプローチによって抽出することを目的としている．本手法では，テキスト中に存在するハイパーリンクを概念としてとらえ，リンクされている部分に着目して解析することで，従来の自然言語処理では精度低下の要因となっていたエンティティ識別の問題を軽減させている．また，関係抽出においては，マルコフモデルに近い条件付き確率モデルの関係ネットワークを構築し，推論ルールの発見に利用している．具体的なデータマイニングの方法としては，まずエンティティ間の関係をデータベースに保存し，「同じクラス（学校）に属する人は友だちであることが多い」といったようなパターンの発見とそのスコアを計算する．最初はデータベースに含まれる関係が少ないため抽出できる推論関係は少ないが，さらにこの処理によって得られた関係を反復して利用することにより，潜在的な推論関係を抽出する．

5.2 カテゴリリンクを利用する方法

Wikipedia から概念間の関係を抽出する研究でよく利用されるデータの 1 つがカテゴリ

ツリーである．前述のとおり Wikipedia のカテゴリツリーは，概念分類に利用されているため，概念の is-a 関係抽出に向いている．しかし，Wikipedia のカテゴリツリーは複数の親ノードへの所属や循環を許す複雑なネットワーク構造をしており，通常のツリー構造ではないため，単にツリー構造をたどって親子関係を抽出するとまったく関係しない概念関係が抽出される可能性がある．また，カテゴリは単なる分類であり，is-a 関係になっていない場合も多い．

そこで，Ponzetto らの研究[22] では，主にカテゴリ名に対して文字列照合を行うことによって，Wikipedia のカテゴリリンクを is-a 関係と not-is-a 関係に分類し，is-a 関係にあるカテゴリ名のペアを収集する手法を提案している．たとえば，カテゴリ名 British Computer Scientists と Computer Scientists は，解析して得られた語彙の主要部分（lexical head）を共有しているので is-a 関係に分類される．ここでは，Computer Scientists が British Computer Scientists の語彙の主要部分である．一方で，Crime Comics と Crime のように，Crime Comics の語彙の主要部分（ここでは Comics）が，もう一方のカテゴリ名の先頭に出現しない場合は，not-is-a 関係に分類される．ResearchCyc [⋆1] と比較するために，カテゴリのペアに含まれる各概念を，ResearchCyc の概念とマッピングしたうえで比較した結果，この手法は 73.7%の recall と 100%の precision が得られている．またテキストコーパスからパターンマッチングによって得られた is-a，not-is-a 関係の情報をもとにカテゴリ名のペアを分類した結果も含めると，84.3%の recall と 91.5%の precision を実現している．具体的には，参考文献 22) の Figure 1 に示される $NP1$ is-a $NP2$ の関係を定義する “$NP2$ like $NP*NP1$”（ex：*stimulants* like *caffeine*）のような英文法に特化した is-a 関係と not-is-a 関係のパターンを定義し，それをテキストコーパスに適用することによって is-a，not-is-a 関係の単語ペア数の統計を収集する．その情報をもとにカテゴリ名のペアを is-a と not-is-a 関係に分類を行う．

一方，YAGO [26),27)] は，Wikipedia と WordNet を利用したオントロジである．YAGO では，オントロジ中のクラスを Wikipedia のカテゴリ名から，そのクラスに属するインスタンスを記事タイトルから収集している．クラスを収集するために，すべての Wikipedia カテゴリから，Conceptual Category と呼ばれるカテゴリを識別しクラスとする．Conceptual Category とは，特に概念の意味的な表現に利用されるカテゴリであり，管理目的のカテゴリ（例：Articles with unsourced statements, Articles with dead external links）や，誕生年などある概念の関連情報を表すカテゴリ（例：1979 births）とは区別される[26]．Conceptual Category は，前述の管理や関連情報を目的とするカテゴリを除いたうえで，英文法に特化したカテゴリ名の言語処理によって識別され，たとえば，記事「Albert Einstein」が属するカテゴリ「Naturalized citizens of the United States」は Conceptual Category である．この Conceptual Category に属する Wikipedia の記事タイトル（例：Albert Einstein）をインスタンスとして収集する．さらにクラスどうしの is-a 関係を定義するために，WordNet の synset に Conceptual Category をマッピングしている．その結果，YAGO では約 14 万ものクラス間の is-a 関係を約 95%の高精度で抽出している．

また，カテゴリツリーの解析方法に近い手法として，Sumida らの「階層的レイアウト」を利用した上位・下位概念の抽出手法がある[28]．階層的レイアウトとは，セクション名と箇条書きの項目リストから構成されるツリー状の語集合のことである．たとえば，「紅茶」という記事の中には「主な紅茶ブランド」という第 1 レベルのセクションが存在し，その中にはさらに「イギリス」や「フランス」といった第 2 レベルのセクションが存在する．そして，その第 2 レベルのセクションの中にはさらに「Lipton」や「Fauchon」といった箇条書き項目が存在する．これらのセクション名や箇条書きは階層構造で表現可能であり，これらの要素から構成される階層を階層的レイアウトと呼ぶ．提案手法では，概念辞書の概念間の上位・下位概念の候補を抽出するために，階層レイアウトに含まれる語の組合せに対して機械学習手法（SVM）に基づくフィルタリングを行っている．SVM に与える素性には，上位語と下位語の距離，末尾の 1 文字が一致するか，属性語であるかなどの情報が利用される．

5.3　インフォボックスを利用する方法

前述のとおり，インフォボックスは概念のプロパティを記述するための特殊構文である．インフォボックスの解析では，テキストを自然言語処理して関係抽出する手法に対して比較的高精度に構造化されたデータを抽出しやすいのが特徴である．これは，前述の PORE [29)] でもインフォボックスから正解集合を作成していることからも分かる．DBPedia [2)] はインフォボックスから構造化データを抽出する研究の代表例であり，抽出した概念関係を RDF 化し，PARQL での検索インタフェースを提供している．このことにより，複雑なクエリを処理可能なオブジェクト検索などを実現している．

また DBpedia は，意味情報に対して URI を付与し相互にリンクすることによって，異なるデータセットを Web 上で接続・共有する方法である「Linked Data」に基づいて，現在 FOAF（Friend Of A Friend）や GeoNames，DBLP などの様々な外部データとの連携

⋆1 http://research.cyc.com

が行われている[2),⋆1]．とりわけ，DBpedia のような Wikipedia から構築された大規模な概念集合は，他の Semantic Web データに対する仲介役を担うことが期待されている．これは，1 つの統一的な大規模オントロジを仲介することによって，Semantic Web の課題の 1 つである乱立するオントロジ間のマッピングが実現できる可能性があるためである．

意味情報の抽出方法としては，Wikipedia 内のインフォボックスから，出現頻度の低い属性をノイズとして排除し，Subject を記事のタイトル，Predicate をインフォボックスの属性，Object をインフォボックスの属性値として RDF トリプルを生成する．英語版 Wikipedia の約 150 万記事（10 GB）を対象に行った結果，DBpedia では約 840 万の RDF トリプルを抽出している．

5.4　問題点と課題

関係抽出に関する研究において，よく利用されるのがインフォボックスとカテゴリである．インフォボックスを利用した関係抽出では，高い精度や検索への応用などが実現できる反面，網羅性や関係の種類が制限されることなどが問題となる．カテゴリを利用する方法においては，カテゴリ間だけを利用すると網羅性が低くなる一方で，WordNet と融合することによって，大規模で高精度な is-a 関係を構築可能となっている．しかし，より多くの is-a 関係を高精度に収集するためには，複雑なカテゴリネットワーク構造を考慮した解析手法や概念関係の判別が必要とされている．

テキストの解析も概念間の関係抽出の網羅性を高めるためには重要な要素であるといえる．ただし，テキストの解析をする研究の多くが，Wikipedia の一部のデータを利用したものであり，Wikipedia 全体に対して自然言語処理を行った研究は少ない．これは，実際に Wikipedia 全体に適用するにはスケーラビリティの高い手法が必要であるといえる．筆者らの調査では，2009 年 8 月の段階で，英語版 Wikipedia には約 23 GB のテキストデータが存在し，これらすべてのテキストを構文解析などの自然言語処理にかけるためには大量の計算機リソースが必要とされる．また，Wikipedia の記事内では省略表現や代名詞が数多く利用されるため，各文に含まれる主題の照応解析が重要な課題となる．

6.　対訳辞書構築（Bilingual Dictionary Construction）

Wikipedia は 270 以上の言語をサポートし，言語間には多数の言語間リンクが存在する．そのため，単に言語間リンクを抽出するだけでも対訳辞書が構築可能であるが，さらに対訳関係を抽出しようという研究が進められている．

本分野では，単に言語リンクから対訳関係を抽出し，アプリケーションに適用するという研究が多い．たとえば，Bouma らの研究[3)] では多言語に対応した質問応答システムを構築するために，Schönhofen らの研究[24)] では多言語に対応した情報検索のために，また Adafre らの研究[1)] では異言語間の意味的に類似したセンテンスを発見するために，Wikipedia の言語リンクを解析して得られた対訳関係を利用している．さらに，Ferrández らの研究[7)] では，言語リンクを解析して固有名詞の翻訳や曖昧性解消を行っている．

一方，言語リンクを対訳関係として抽出する研究のほかに，言語リンクで結ばれたページどうしが多くの場合同じ内容を示していることを利用し，コンパラブルコーパス（comparable corpus）を構築する試みも行われてきた．Haghighi らは，コンパラブルコーパスを Wikipedia から抽出する手法を提案し[9)]，実験によりその有効性を示している．

上記のとおり，本分野の多くは言語リンクを対訳関係として利用することが主流だが，Erdmann らは言語リンクで網羅されている対訳関係の数に限りがあることを指摘し，言語リンク以外の半構造化情報を利用することで対訳関係が拡張できることを示した[6)]．具体的には，アンカーテキストとリダイレクトリンクを解析することで同義語が抽出できることと，バックワードリンク数などの対訳の信頼性を数値化するために有効である特徴を利用し，教師あり学習（supervised learning）を用いた対訳関係を抽出するアルゴリズムを提案している．

7.　アプリケーション

Wikipedia の解析結果を自然言語処理タスクやアプリケーションに適用するという研究も進められている．この分野には，語義曖昧性解消や情報検索への応用などがあげられる．以降，主な研究領域について解説する．

7.1　語義曖昧性解消への適用

語義曖昧性解消とは，多義性を持つ語が文中で使われているときに，どの意味で使われているかを識別するタスクであり，自然言語処理において重要な研究領域の 1 つである．一般的な語義曖昧性解消は，語の前後に出現する単語やその関係などを手がかりに，辞書やシソーラスを用いて，コンテキストを判定し，最も妥当と推測される語義を決定する．

語義曖昧性解消のための手法で一般的なのは，大量のコーパスから統計的に語義情報を学習するアプローチである．このアプローチでは，特定の語が出現する文章を統計的に解析し，前後に出現する語や文章中に出現する語などを利用して，コンテキストを学習する．こ

⋆1 http://wiki.dbpedia.org/Interlinking

のとき，学習用のコーパスとして語の意味がタグとして明確に付与されたコーパスが存在する教師あり学習手法は，高い精度で語義曖昧性解消が行えるという結果が得られている[23]．しかし，このような語義（タグ）付き学習用コーパスは人手によって作成するために多量のコストを要する点が本アプローチのボトルネックであった．

一方，前述のとおり Wikipedia では URL によって概念が一意に特定され，各ページ（概念）どうしがリンクによって結ばれている．つまり，Wikipedia の文章中に含まれるリンクは語義（タグ）付きの単語と見なすことができる．つまり，文章中でアンカーテキストとして曖昧な語が利用されていたとしても，リンク先としてはコンテキストに沿って適切な概念が参照されている．たとえば，曖昧な語「spring」がアンカーテキストになっている「A spring is an elastic object used to store mechanical energy」という一文を考えると，この spring は機械部品の意味であり，「Spring_(device)」にリンクが張られている．このように，Wikipedia ではアンカーテキストが曖昧な語であるとき，ハイパーリンクを語義（タグ）と見なすことができる．

Mihalcea[15] は，Wikipedia が語義付きの学習コーパスとして利用できることを最初に実証した研究者の 1 人である．上述のように Wikipedia の記事中のリンクを語義タグとして扱い，曖昧性解消のタスクに適用している．また，Mihalcea は，得られた語義を WordNet に手動でマッピングすることで，従来の評価セットでの評価が行える語義タグ付きコーパスを作成した．構築したコーパスの有用性を示すために，コンテキスト情報を単純ベイズ分類器で学習し，SENSEVAL-2 と SENSEVAL-3 ⋆1のテストコレクションで評価した．結果として，Wikipedia から得られたコーパスで学習した分類器は，2 つの単純なベースライン手法である MFS ⋆2と Lesk-corpus ⋆3に比べて語義曖昧性解消の相対誤差が 30〜44%減少した．しかし，SENSEVAL の訓練コーパスとの比較から，Wikipedia によって識別される語義は全体的に WordNet で定義されている語義に比べて数が少なく，一般的に粗いことが述べられている．さらに，Wikipedia で定義されているほとんどが名詞である点や，語義の頻度分布が偏っている点，ときおりハイパーリンクのリンク先が間違っている点などが問題としてあげられているが，それらを差し引いても十分な量の語義データが得られるという長所は利用価値の高いものであると Mihalcea は結論づけている．

さらに Mihalcea ら[14] は，この語義曖昧性解消を利用したアプリケーションとして，与えられた文書に対して Wikipedia の記事へのリンクを付与するアプリケーション Wikify!を実装している．本タスクは，文書からキーワードを抽出し，それを曖昧性解消してリンク先の概念を識別することで行われる．キーワード抽出は，Wikipedia の記事中にアンカーテキストとして出現した語すべてを N-gram を用いて与えられた文書から抽出し，それらを候補としてキーワードの判定を行う．判定のアルゴリズムとしては，tf-idf や Keyphraseness（語の出現回数を利用するスコアリング）などの手法を適用し，比較評価した結果，Keyphraseness が精度・再現率ともに高い値を示した．本タスクは情報検索分野の INEX 評価型会議において Wikipedia を題材とするリンク発見タスク⋆4として行われている．

Bunescu ら[4] は，Wikipedia を用いた固有表現を抽出する手法と固有表現に対して曖昧性解消を行う手法を提案している．本手法では，与えられた単語が固有表現であるかを判定するために，記事タイトルや文中での大文字・小文字の使われ方を少数の規則として作成し，それを用いて固有表現を識別する．たとえば，記事タイトルが EU であった場合，2 文字目も大文字であるという情報から固有表現である可能性が高いことや，逆に記事中で利用されるときに頭字語が小文字として記述されていた場合は，固有表現である可能性が低いと考えられるなどといったルールで固有表現であるかを判別する．Bunescu らは，語義曖昧性解消のベースライン手法として，曖昧な語が出現するコンテキスト情報を，曖昧な語を中心にその前後 55 語で表現する手法を提案している．この手法では，コンテキスト情報と語義をそれぞれ TF-IDF で重み付けした特徴ベクトルで表現し，それらのコサイン類似度が最大となる概念を採用するアプローチである．しかし，この手法では，記事が短すぎる場合や，不完全である場合，同じことを述べているが表記が異なる場合に類似度が計算できないという問題がある．この問題を解決する手法として，前後に出現する 55 語と候補となる概念のカテゴリとの相関を用いた曖昧性解消手法を提案している．語とカテゴリの相関は，Wikipedia から得られた語義タグ付きコーパスを学習データとして SVM で学習している．学習のための素性は，曖昧な語の前後の 55 語と語義のカテゴリ（上位カテゴリも利用）との共起を用いている．すべてのカテゴリを用いると極端に時間がかかるので，実際に本手法を用いるときに使われるであろう 4 種類のカテゴリ選択に関して提案手法を評価した．結果として，4 種類のカテゴリ選択すべてにおいて提案手法のほうがベースライン手法に比べて高い精度を実現することができた．

上述のとおり，Wikipedia は，URL によって語義が一意に特定できる点や，曖昧性を解

⋆1 http://www.senseval.org/
⋆2 MFS (most frequent sense)：語義タグ付きコーパスで最も出現した語義に曖昧性解消する手法
⋆3 Lesk-corpus：コーパスをもとに Lesk アルゴリズムによって曖昧性解消を行う手法[11]
⋆4 http://www.inex.otago.ac.nz/tracks/wiki-link/wiki-link.asp

消するためのページも存在することから，語義曖昧性解消のアルゴリズムを構築する際のデータとして有効であることが証明されてきた．しかしその一方で，Wikipedia で定義されているほとんどが名詞であることから，その他の品詞表現への対応は 1 つの課題となる．また，Wikipedia に出現する語義（タグ）は分野によって出現する頻度が偏っていることも多いため，この偏りを考慮（補間）する手法も研究の余地がある．さらに不特定多数のユーザが書き込むことから，コンテンツ自体に間違いがあることも少なくなく，精度にも影響する場合がある．たとえば，ある概念を記述するページが複数存在する誤りは多くみられ，これは曖昧性解消の精度に大きく影響する．当然，それらの問題を差し引いても，大量の学習コーパスを得ることができるメリットは大きいといえるが，さらなる精度向上の余地はあるといえる．

7.2　情報検索への応用

Milne ら[16] は，Wikipedia から抽出した知識を利用した新しい情報検索インタフェース Koru を実装している．Koru の情報検索インタフェースは，複数のモジュールから構成される．まず，1 つ目のモジュールは，クエリからトピックを抽出し，情報検索を支援する機能を提供する．たとえば，ユーザが「american airline」とクエリを入力した場合，Koru は，「American Airlines」，「Airline」，「Americas」などをトピックとしてユーザに提示する．ユーザはこれらのトピックの中から任意のものを選択し，情報検索の結果を絞り込むことができる．ここで重要なポイントは，各トピックは Wikipedia のページに対応付けられるという点である．この結果，語の曖昧性が解消された状態で情報の絞り込みが可能となり，よりユーザの意図に沿った情報検索が可能になる．たとえば，先ほどの例では，「American Airlines」を選択すれば固有名詞としての「American Airlines」に関係する情報だけを絞り込むことができ，「Airline」と「Americas」を選択した場合，アメリカと航空会社に関係する情報だけに絞り込んで情報検索することが可能である．

2 つ目のモジュールは，意味の近い単語を利用したクエリ拡張機能を提供する．本機能は，Wikipedia のリダイレクトページを用いて実現されている．これは，リダイレクトページがリダイレクト先のページに意味的に近い語（たとえば同義語や下位語）を定義するためによく利用されることに着目した手法である．さらに，表示されているトピックに対して関係の強いトピックをリスト形式で表示させるモジュールが存在し，クエリ拡張を可能にしている．このモジュールは，関連の強い概念を抽出するために，2 つの記事内に含まれるリンクを比較し，関連度を計算している．ほかにも，概念間の関連度を利用してトピックの語義曖昧性解消をするモジュールやトピックの重要度を利用して提示されたトピックから重要なものを自動的に選択するモジュールが存在する．

Koru は，従来のキーワード検索と比較して情報検索の精度を落とすことなく再現率を大きく向上させている．Milne は，この大きな再現率向上から Wikipedia のリダイレクトページの質が高いことを証明しており，精度が落ちなかったのは自動抽出された句によるものだと考えられると述べており，情報検索での Wikipedia の有用性を示した．

8. ま と め

本稿では，現状の Wikipedia 研究を俯瞰するとともに，その問題点を議論した．**表 1** に本稿で解説した論文をまとめた．全体としては，連想関係抽出に関してはカテゴリツリーの解析，ページ間リンクの解析，コンテンツ類似性などの種々の指標が出揃った感がある．また，各種の研究用フレームワークや抽出されたデータセットなども公開されており，それらのツールを利用した研究も進められている．以下に代表的なプロジェクトを示す．

- JWPL（http://www.ukp.tu-darmstadt.de/software/jwpl/）
  Wikipedia のリンク情報やカテゴリ情報などの各種構造データにアクセスできる Java のライブラリ．
- DBPedia（http://dbpedia.org/）
  Wikipedia から抽出した構造化データへの各種クエリ処理機構を提供している．URI ベースのクエリから，SPARQL を利用したクエリまで様々なインタフェースを提供している．
- Wikipedia-lab（http://wikipedia-lab.org/）
  リンクテキストやセクション情報，インフォボックスなどの情報をデータベースに格納し，インデックスを付与するスクリプトである「Wikipedia RR（Research Resources）」や，Wikipedia から抽出した連想辞書に対する API（Wikipedia API）などを公開している．

表 1　研究の分類
Table 1　Research classification.

リソース	関連度計算	関係抽出	対訳辞書構築	曖昧性解消	情報検索	その他
カテゴリリンク	25)	22)，26)–28)		4)		21)
ページ間リンク	10)，17)，19)		6)	4)，14)，15)	16)	
テキスト	8)，25)	5)，18)，29)		4)		21)
インフォボックス		2)				
言語間リンク		1)	3)，7)	7)	24)	9)

- Wik-IE（http://wik-ie.sourceforge.jp/）
  記事やカテゴリ・リダイレクト間の関係や他言語版へのリンク情報などを抽出するツール「Wik-IE」を公開している．Hadoop を利用した分散処理によって高速にダンプデータを解析できるのが特徴．

関係抽出やオントロジの分野では，ブートストラッピング手法による関係の抽出が主流になりつつあるが，まだアプローチも出揃った感はなく，改良の余地も大いにあると思われる．言語間リンクの研究に関しては，単に言語間リンクを抽出して何らかのタスクによって評価した，といったレベルの研究が多く，まだ様々な指標やリソースを組み合わせた手法が研究されていくことが予想される．

さらに，Wikipedia マイニングの応用としては，語の曖昧性回避など自然言語処理の基礎的な部分への適用は十分に進められているものの，さらに多くのアプリケーションによって Wikipedia マイニングの有用性が示されることが必要であると思われる．

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と，科学研究費補助金基盤研究（C）（20500093）および科学研究費補助金基盤研究（B）（21300032）によるものである．ここに記して謝意を表す．

参 考 文 献

1) Adafre, S.F. and de Rijke, M.: Finding Similar Sentences across Multiple Languages in Wikipedia, *Proc. EACL Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, pp.62–69 (2006).
2) Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data, *Proc. International Semantic Web Conference, Asian Semantic Web Conference* (*ISWC/ASWC*), pp.722–735 (2007).
3) Bouma, G., Fahmi, I., Mur, J., van Noord, G., vander Plas, L. and Tiedemann, J.: The University of Groningen at QA@CLEF 2006 Using Syntactic Knowledge for QA, *Working Notes for the Cross Language Evaluation Forum Workshop* (*CLEF*) (2006).
4) Bunescu, R.C. and Pasca, M.: Using Encyclopedic Knowledge for Named entity Disambiguation, *Proc. European Chapter of the Association for Computational Linguistics* (*EACL*) (2006).
5) Culotta, A., McCallum, A. and Betz, J.: Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text, *Proc. Human Language Technology Conference — North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*) (2006).
6) Erdmann, M., Nakayama, K., Hara, T. and Nishio, S.: Bilingual Dictionary Extraction from Wikipedia Using an SVM Classifier, *ACM Trans. Multimedia Computing, Communications and Applications* (*ACM TOMCCAP*) (2009).
7) Ferrández, S., Toral, A., Ferrández, Ó., Ferrández, A. and Noz, R.M.: Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering, *Proc. International Conference on Applications of Natural Language Processing and Information Systems* (*NLDB*), pp.352–363 (2007).
8) Gabrilovich, E. and Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, *Proc. International Joint Conference on Artificial Intelligence* (*IJCAI*), pp.1606–1611 (2007).
9) Haghighi, A., Liang, P., Berg-Kirkpatrick, T. and Klein, D.: Learning Bilingual Lexicons from Monolingual Corpora, *Proc. Association for Computational Linguistics* (*ACL*) (2008).
10) Ito, M., Nakayama, K., Hara, T. and Nishio, S.: Association Thesaurus Construction Methods based on Link Co-occurrence Analysis For Wikipedia, *Proc. ACM International Conference on Information and Knowledge Management* (*CIKM*), pp.817–826 (2008).
11) Kilgarriff, A. and Rosenzweig, J.: Framework and results for English SENSEVAL, *Computers and the Humanities*, Vol.34, No.1, pp.15–48 (2000).
12) Leuf, B. and Cunningham, W.: *The Wiki Way: Collaboration and Sharing on the Internet*, Addison-Wesley (2001).
13) Li, X. and Liu, B.: Learning to Classify Texts Using Positive and Unlabeled Data, *Proc. International Joint Conference on Artificial Intelligence* (*IJCAI*), pp.587–594 (2003).
14) Mihalcea, R. and Csomai, A.: Wikify!: linking documents to encyclopedic knowledge, *Proc. 16th ACM Conference on information and knowledge management*, New York, NY, USA, pp.233–242, ACM (2007).
15) Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation, *Proc. North American chapter of the Association for Computational Linguistics* (*NAACL*) (2007).
16) Milne, D.N., Witten, I.H. and Nichols, D.M.: A Knowledge-based Search Engine Powered by Wikipedia, *Proc. ACM International Conference on Information and Knowledge Management* (*CIKM*), pp.445–454 (2007).
17) Nakayama, K., Hara, T. and Nishio, S.: Wikipedia Mining for An Association Web Thesaurus Construction, *Proc. International Conference on Web Information Systems Engineering* (*WISE*), pp.322–334 (2007).
18) Nguyen, D.P.T., Matsuo, Y. and Ishizuka, M.: Relation Extraction from Wikipedia Using Subtree Mining, *Proc. Conference on Artificial Intelligence* (*AAAI*),

pp.1414–1420 (2007).
19) Ollivier, Y. and Senellart, P.: Finding Related Pages Using Green Measures: An Illustration with Wikipedia, *Proc. Conference on Artificial Intelligence* (*AAAI*), pp.1427–1433 (2007).
20) Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proc. International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics* (*COLING/ACL*) (2006).
21) Ponzetto, S. and Strube, M.: An API for Measuring the Relatedness of Words in Wikipedia, *Proc. Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.49–52 (2007).
22) Ponzetto, S. and Strube, M.: Deriving a Large Scale Taxonomy from Wikipedia, *Proc. Conference on Artificial Intelligence* (*AAAI*), pp.1440–1447 (2007).
23) Pradhan, S., Loper, E., Dligach, D. and Palmer, M.: Semeval-2007 task-17: English lexical sample, srl and all words, *Proc. 4th International Workshop on Semantic Evaluations* (*SemEval-2007*), pp.87–92 (2007).
24) Schönhofen, P., Benczúr, A., Bíró, I. and Csalogány, K.: Performing Cross-Language Retrieval with Wikipedia, *Working Notes for the Cross Language Evaluation Forum Workshop* (*CLEF*) (2007).
25) Strube, M. and Ponzetto, S.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, *Proc. Conference on Artificial Intelligence* (*AAAI*), pp.1419–1424 (2006).
26) Suchanek, F.M., Kasneci, G. and Weikum, G.: YAGO: A Core of Semantic Knowledge, *Proc. International Conference on World Wide Web* (*WWW*), pp.697–706 (2007).
27) Suchanek, F.M., Kasneci, G. and Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet, *Journal of Web Semantics*, Vol.6, No.3, pp.203–217 (2008).
28) Sumida, A., Yoshinaga, N. and Torisawa, K.: Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia, *Proc. Language Resources and Evaluation Conference* (*LREC*) (2008).
29) Wang, G., Yu, Y. and Zhu, H.: PORE: Positive-Only Relation Extraction from Wikipedia Text, *Proc. International Semantic Web Conference, Asian Semantic Web Conference* (*ISWC/ASWC*), pp.580–594 (2007).
30) 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（アーティクル），人工知能学会誌，Vol.22, No.5, pp.693–701 (2007).

(平成 21 年 6 月 20 日受付)
(平成 21 年 10 月 6 日採録)

（担当編集委員　市川 哲彦）

中山浩太郎（正会員）

2001 年関西大学総合情報学部卒業．2003 年同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社長，同志社女子大学非常勤講師に就任．2004 年関西大学大学院を中退後，2007 年大阪大学大学院情報科学研究科にて博士号を取得し，同年 4 月大阪大学大学院情報科学研究科特任研究員，2008 年 4 月東京大学知の構造化センター特任助教に就任し，現在に至る．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．IEEE，ACM，電子情報通信学会，人工知能学会の各会員．

伊藤　雅弘（学生会員）

2007 年立命館大学理工学部情報学科卒業．2008 年大阪大学大学院情報科学研究科博士前期課程修了．同年 4 月同大学院情報科学研究科博士後期課程に進学，現在に至る．人工知能，WWW からの知識獲得および情報検索に関する研究に興味を持つ．日本データベース学会，人工知能学会の各学生会員．

Erdmann Maike

2006 年ドイツ・CvO 大学卒業．2008 年大阪大学大学院情報科学研究科博士前期課程修了．同年 4 月同大学院情報科学研究科博士後期課程に進学，現在に至る．自然言語処理，WWW からの知識獲得に関する研究に興味を持つ．

白川　真澄（学生会員）

2008 年大阪大学工学部電子情報エネルギー工学科卒業．同年 4 月同大学院情報科学研究科博士前期課程に進学，現在に至る．人工知能，WWW からの知識獲得および情報検索に関する研究に興味を持つ．日本データベース学会の学生会員．

道下　智之

2008 年神戸大学情報知能工学科卒業．同年 4 月大阪大学大学院情報科学研究科博士前期課程に進学，現在に至る．人工知能，WWW からの知識獲得および情報検索に関する研究に興味を持つ．

原　　隆浩（正会員）

1995 年大阪大学工学部情報システム工学科卒業．1997 年同大学院工学研究科博士前期課程修了．同年同大学院工学研究科博士後期課程中退後，同大学院工学研究科情報システム工学専攻助手，2002 年同大学院情報科学研究科マルチメディア工学専攻助手，2004 年同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．1996 年本学会山下記念研究賞受賞．2000 年電気通信普及財団テレコムシステム技術賞受賞．2003 年本学会研究開発奨励賞受賞．2008 年，2009 年情報処理学会論文賞．データベースシステム，分散処理に興味を持つ．IEEE，ACM，電子情報通信学会，日本データベース学会の各会員．

西尾章治郎（正会員）

1975 年京都大学工学部数理工学科卒業．1980 年同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情報システム工学専攻教授を経て，2002 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，その後 2007 年より大阪大学理事・副学長に就任．この間，カナダ・ウォータールー大学，ビクトリア大学客員．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．本会理事を歴任．本会論文賞を受賞．電子情報通信学会フェローを含め，ACM，IEEE 等 8 学会の各会員．