複数語句から構成されるコンテキストを考慮した連想関係の抽出

白川　真澄† 　中山浩太郎†† 　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

あらまし　連想関係の抽出，すなわち与えられた語に対して関連している語を取得する技術は，連想検索やクエリフリー検索，文書分類など，テキスト情報から意味を推測する処理を含むアプリケーション全般において重要な基盤技術である．筆者らは先行研究において，Wikipedia を解析することにより，エンティティ間の関連度を定義した大規模な連想辞書を構築してきた．複数語から連想される語は，それら語句が構成するコンテキストに依存するが，従来研究ではコンテキストの取り扱いが技術的課題であった．本研究では，複数語句が入力として与えられたときに，それらの語句が構成するコンテキストを考慮した上で，連想されるエンティティを抽出する手法を提案する．

キーワード　Wikipedia マイニング, 連想辞書, コンテキスト依存，ブートストラッピング法

Extraction of Association Relations from Contexts Consisting of Multi-words

Masumi SHIRAKAWA†, Kotaro NAKAYAMA††, 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

1. 研 究 背 景

Web は，人々が情報を取得するための手段として急速に普及し，今や我々の生活に欠かせない存在となっている．全世界における 1ヶ月の Web 検索クエリの発行数は 610 億にも上ると報告されている（2007 年 8 月，comScore(注1) qSerach 2.0）．しかし，Web の情報の多くは単なるテキストとして表現されており，意味を考慮した情報検索の実現が大きな課題となっている．そのため，ユーザが自分の欲しい情報を Web から取得しようとした場合，ユーザ自身が発行した Web 検索クエリに対して単純なキーワード検索を用いるのが主流となっている [1]．

意味を考慮した情報検索を実現するための研究としては，クエリ拡張 [2] やクエリフリー検索 [3]，Web ページの分類 [4] など様々なものがあるが，その基盤技術の一つとして，ある語から連想される語（すなわち連想関係）を取得する技術が挙げられる．連想関係を定義した辞書である連想辞書は，上記のアプリケーションをはじめ，意味を考慮した情報検索を実現する様々なアプリケーションの基盤技術として利用可能である [5], [6]．

筆者らはこれまで，連想関係を取得する研究として，Wikipedia シソーラス(注2)と呼ばれる大規模な連想辞書を構築してきた [7], [8]．Wikipedia シソーラスは，Wikipedia で定義されているエンティティ（記事）間の関連度を定義した連想辞書であり，語を一つ入力すると，その語から連想されるエンティティ集合がスコア付きで出力される．

しかし，入力として複数の語を想定した場合，それらの語はコンテキストを構成するため [3]，連想されるエンティティ（あ

(注1)：http://www.comscore.com/

(注2)：http://dev.sigwp.org/WikipediaThesaurusV3/

るいは語）もそのコンテキストに応じて変化すると考えられる．例えば,「富士山」と「阿蘇山」という語の組合せからは山に関するエンティティが連想されやすく，また,「富士山」と「熱海」という語の組合せからは静岡に関するエンティティが連想されやすくなる．Wikipedia シソーラスでは，一つの語句に対して連想されるエンティティ集合を出力するが，入力が複数語句の場合，上記の例のようにコンテキストを考慮した上で，連想されるエンティティの集合を取得する必要がある．

そこで本研究では，Wikipedia シソーラスを拡張し，複数の語が入力として与えられたときに，その語句群から構成されるコンテキストを推測し，そのコンテキストに依存した連想関係を抽出する手法 (Wikipedia Sets) を提案する．具体的には，複数の入力語に対し，それぞれ連想されるエンティティとその関連度を取得した後，共通のエンティティの関連度をマージして出力する．また，得られた出力を再入力するというプロセスを反復すること（ブートストラッピング法）により，入力語が構成するコンテキストを強調させ，出力の精度及び網羅性向上を図る．本手法は，前処理が不要でかつ外部の情報を使用しないため，(1) 簡単に実装でき，(2) 拡張性が高く，(3) Wikipedia シソーラス以外の連想辞書にも適用可能である．

2. 関 連 研 究

連想辞書は，与えられた語句から連想される語（あるいはエンティティ）を取得する辞書であり，自然言語処理や情報検索など，幅広い研究領域で利用されてきた [5], [6]．連想辞書の主な構築手法として，自然言語処理を用いた手法 [5], [6], [9], [10] と Web のリンク構造を用いた手法 [11] が挙げられる．自然言語処理を用いた連想辞書構築に関する研究の歴史は古く，コーパス解析により（半）自動的に構築する手法が数多く提案されてきた．例えば，語の共起に基づく手法 [5], [6] やクラスタリングを用いた手法 [10]，語のフィルタリングを用いた手法 [9] などがある．Web のリンク構造を用いた手法では，リンクで繋がっている Web ページは同じトピックについて記述されていることが多いという特性を利用する．また一般に，自然言語処理を組み合わせて用いる場合が多い．例えば Chen らの研究 [11] では，Web サイトの階層構造からサブツリーと呼ばれる木を形成し，木の中の語の共起性を利用して語句間の関連度を計測することで連想辞書を構築している．また，筆者らの先行研究では，Wikipedia を用いて連想辞書を構築している [7], [8]（次章で詳述する）．

連想辞書は，語句間（あるいはエンティティ間）の関連度 (semantic relatedness) を定義したものであるが，これは類似度 (semantic similarity) とは異なる．具体的には，関連度は二つの語が何らかのコンテキストを共有している場合に高い値が与えられるが，類似度はより具体的で，二つの語がどの程度似ているかを示す指標である [12]．一つの語あるいは複数語の入力に対して，類似している語を取得する研究としては，Google Sets(注3)や Bayesian Sets [13]，SEAL [14](注4)などが挙げられる．これらの研究では，特に複数語の入力に対して，類似した語を取得することに重点を置いている．すなわち，入力語句群に対し，最も特徴的な共通点を見つけ，その共通点において類似した語を取得するよう設計されている．

一方，複数語の入力に対して関連している語（あるいはエンティティ）を取得する研究としては，前述の自然言語処理を用いた連想辞書の構築手法 [6] がある．文献 [6] では，関連文書検索の技術を応用して，文書中の語を出力とすることで，自然文クエリ（複数語からなるクエリ）に対して関連語を取得している．また，クエリ拡張（クエリ推薦）に関する研究の中には，複数語の入力に対して関連している語を抽出する研究としてみなせるものも存在する．例えば，Preferred Infrastructure 社の連想検索エンジン reflexa(注5)は，ESA [15] の関連度計算の手法（Wikipedia に出現する語を Wikipedia の記事でベクトル化する方法）を応用させた手法によって，複数語からなるクエリに対しても，コンテキストを考慮したクエリ推薦を行っている(注6)．また，TermCloud [16] は，複数語からなる Web 検索クエリに対して，Web 検索結果によく出現する語句をコンテキスト依存の関連語として抽出し視覚化する．

3. 先 行 研 究

3.1 Wikipedia シソーラス

筆者らはこれまで，Wikipedia を用いて大規模な連想辞書である Wikipedia シソーラス [7], [8] を構築してきた．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，一般的なエンティティから新しいエンティティに至るまで記事が網羅されており，記事（エンティティ）数は，最も多い英語版で 350 万記事，日本語版で 70 万記事である（2011 年 1 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性質を数多く持っている [17]．筆者らは，このような Wikipedia の性質を活かし，Wikipedia のリンク構造を解析することで Wikipedia シソーラスを構築し，Web 上で公開している．2011 年 1 月現在，Wikipedia シソーラスは英語版で 3 億 8000 万以上，日本語版で 1 億以上のエンティティペアに対して関連度を定義している．また，精度の面でも，既存の共起性解析を用いた手法 [5] や Web のリンク構造を用いた手法 [11] と比較して高い精度を達成していることが先行研究において実証されている [7].

Wikipedia シソーラスは，図 1 に示すように，一つの語を入力とし，その語から連想されるエンティティ集合をスコア（関連度）付きのリストとして出力する連想辞書である．Wikipedia シソーラスは，Wikipedia のリンクテキストから抽出したラベ

(注3)：http://labs.google.com/sets
(注4)：http://www.boowa.com/
(注5)：http://labs.preferred.jp/reflexa/
(注6)：アルゴリズムの詳細は公開されていない.

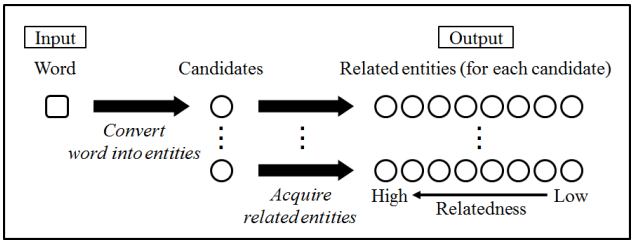

図 1　Wikipedia シソーラス
Fig. 1　Wikipedia Thesaurus

ル情報 [18] を語とエンティティの変換テーブルとして保持しているため，入力語が多義語である場合も，それぞれの意味に対して個別に関連エンティティを出力可能である．

3.2　課　　題

Wikipedia シソーラスの課題として，複数の入力への対応が挙げられる．入力が複数語句である場合，それらの語句が構成するコンテキストを考慮した上で連想関係を取得するべきである．例えば,「富士山」と「阿蘇山」という語の組合せは，山というコンテキストを構成しており，山に関するエンティティが連想されやすくなる．一方,「富士山」と「熱海」という語の組合せの場合，これらの語句は静岡県というコンテキストを構成しているため，静岡県に関するエンティティが連想されやすくなる．このように，複数の入力語がどのようなコンテキストを構成しているかによって，それらの語から連想されるエンティティも異なってくる．同じコンテキストを共有しているエンティティ同士は同時に連想されやすい一方，同じコンテキストを共有していないエンティティ同士が同時に連想されることは少ない．

4.　Wikipedia Sets

4.1　コンテキストを考慮した関連エンティティの取得

本研究では，複数語句を入力として与えたときに，それら入力のコンテキストに沿って連想されるエンティティを抽出する手法 (Wikipedia Sets) を提案する．

先行研究である Wikipedia シソーラスはエンティティ間の関連度を定義しており，語を入力すると，Wikipedia のリンクテキストから抽出したラベル情報 [18] を利用して，語をエンティティに変換してから関連エンティティを取得する．入力語が多義語である場合は，それぞれの意味のエンティティに対して別々に関連するエンティティ集合を出力する．ここで，入力が複数語である場合を考えると，入力語が構成するコンテキストに依存した関連エンティティを発見するために，各入力語に対して得られた関連エンティティをどのように処理すればよいかという技術的な問題が発生する．

この問題に対して，本手法では，Gabrilovich らの手法 [15] で用いられている語義曖昧性解消の考え方を採用し，入力語句群に対して共通して出現する関連エンティティを優先的に出力する．本手法は，前処理が不要でかつ外部の情報を使用しないため，(1) 簡単に実装でき，(2) 拡張性が高く，(3) Wikipedia シソーラス以外の連想辞書にも適用可能である．Gabrilovich らの語義曖昧性解消の考え方とは，複数の語は，それぞれ他の語同士で意味情報を補い合うことにより，語義を決定するコンテキストが強調されるため，語義曖昧性解消が可能である，というものである．本手法においても，複数語からなる入力に対し，それぞれ他の入力語によって語義を強め合うことにより語義曖昧性解消を行う．また，この考え方を拡張し，複数語からなる入力に対して他の入力語と共通しているコンテキストの推測を試みる．ただし，この手法において得られるのは，入力語に対して語義曖昧性解消およびコンテキスト推定を行ったときの出力（関連エンティティ）であり，入力語がどのような語義であったか，どのようなコンテキストに依存して出力が得られたかについては，得られた出力からさらに何らかの手法を用いて推定する必要がある．

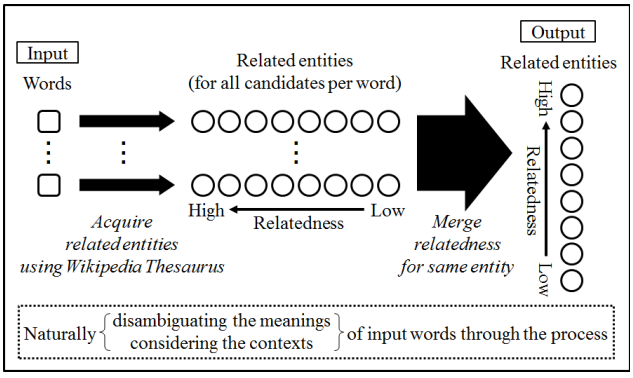

図 2　Wikipedia Sets: Wikipedia シソーラスを用いた複数語句からの関連エンティティ取得
Fig. 2　Wikipedia Sets: Acquisition of related entities from multi-words using Wikipedia Thesaurus

本手法のモデルを図 2 に示す．まず，各入力語に対して，それぞれ Wikipedia シソーラスを用いて関連するエンティティおよびその関連度を取得する．ただし，Wikipedia シソーラスはエンティティ間の関連度を定義した連想辞書であるため，図 1 と同様に，各入力語に対して，それぞれ意味しうる全てのエンティティ（語義）に変換した後，各エンティティについて関連エンティティのリストを取得する．図 1 と異なるのは，各語義について別々に関連エンティティを取得した後，入力語ごとにそれらを一つのリストに集約する点である．各入力語ごとに関連するエンティティを抽出した後，同じ関連エンティティについて関連度をマージすることで，入力語句群に対する最終的な関連エンティティと関連度のリストを得る．最終的な関連度は，入力語が構成するコンテキストにどの程度関連しているかを意味している．なお，本手法の実装においては，関連度のマージ方法として単純に加算する方法を用いている．関連度のマージ方法については様々な方法が考えられるが，使用する連想辞書の特性によって最適な方法が異なると考えられる．

4.2　手法の実装とアルゴリズム

前節で述べた手法 (Wikipedia Sets) を実装し，与えられた入力語句群に対してコンテキストに依存した関連エンティティセットを出力するシステムを構築した．

擬似コードによる Wikipedia Sets のアルゴリズムを Algo-

Algorithm 1 Pseudo-code of Wikipedia Sets

Input: words W

Output: associated entities and their relatedness $E[]$

1: $E[] \leftarrow \phi$ //Initialize by null vector
2: **for** word $w \in W$ **do**
3: $R[] \leftarrow \phi$
4: $C[] \leftarrow$ ConvertWordIntoEntities(w) //With confidence
5: **for** entity $c \in C$.keys **do**
6: $R_{tmp}[] \leftarrow$ AcquireRelatedEntities(c) //With relatedness
7: **for** entity $r \in R_{tmp}$.keys **do**
8: $R[r] \leftarrow \max(R[r], C[c] * R_{tmp}[r])$
9: **end for**
10: **end for**
11: **for** entity $r \in R$.keys **do**
12: $E[r] \leftarrow E[r] + R[r] + Reward$
13: **end for**
14: **end for**

rithm 1 に示す．本アルゴリズムでは，語句集合 W を入力とし，関連エンティティと関連度の連想配列 $E[]$ を出力とする．なお，連想配列 $C[]$，$R[]$，$R_{tmp}[]$ はそれぞれ，語が意味するエンティティ候補，各語の関連エンティティ，各語が意味するエンティティ候補の関連エンティティを，スコア付きで保持するために用いる．まず 4 行目において，それぞれの入力語に対して，それが意味しうるエンティティ候補を信頼度付きで取得している．なお，ここでは Wikipedia のリンクテキストから抽出したラベル情報 [18] を用いている．ラベルとしての信頼度は，様々なものが考えられるが，ここでは文献 [18] で用いられている CS 値の分母と分子をそれぞれ自然対数で正規化した値，すなわち下記のものを使用した．

$$CS_{Normalize}(v_i, q) = \frac{\ln(Cnt(B_{v_i}|q))}{\ln(\sum_{v_j \in V} Cnt(B_{v_j}|q))} \quad (1)$$

$Cnt(B_{v_i}|q)$ は Wikipedia の記事（エンティティ）v_i にリンクテキスト q としてリンクが張られている回数である．なお，$CS_{Normalize}$ の最大値は 1 である．次に，6 行目では，4 行目で取得した全てのエンティティ候補に対して，Wikipedia シソーラスを用いて関連するエンティティとその関連度（0 から 1 までの連続値）を取得している．8 行目でその関連エンティティを，信頼度 $CS_{Normalize}$ の重み付きで保存しているが，同じラベルを持つ別のエンティティが，同じ関連エンティティを持っていた場合，それらのうちで高いほうの関連度を保存する．最後に 12 行目で，各入力語から連想されるエンティティについて，関連度を加算する．ここで，関連度を加算するごとに報酬として $Reward$ を足すことにより，より多くの入力語から共通して連想されるエンティティが優先的に出力されるようになる．関連度の最大値が 1 であるため，Wikipedia Sets では $Reward$ を固定値 1 としている．なお，本システムの関連度のマージ方法はヒューリスティックにより決定しており，今後どのような方法が適切かを検証する必要がある．

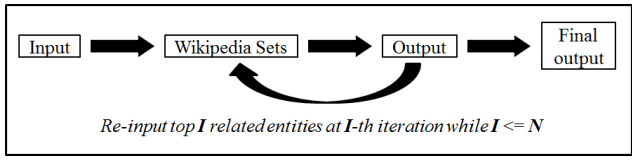

図 3　Wikipedia Sets にブートストラッピング法を適用した例
Fig. 3　Wikipedia Sets with Bootstrapping

4.3　連想関係のフィードバック（ブートストラッピング法）

4.1 節で述べた手法において，出力の関連エンティティの関連度は，入力語が構成するコンテキストにどの程度関連しているかを表している．そこで，出力の上位の関連エンティティが，入力のコンテキストに沿って連想されるべきエンティティとして正しいものと仮定し，それらのエンティティをフィードバック（再入力）させる手法を提案する．得られた出力を再入力するというプロセスを反復すること（ブートストラッピング法）により，入力の数を擬似的に増やすことができる．その結果，入力語句群が構成するコンテキストが補強され，コンテキストを推測しやすくなり，より安定した出力を取得できるようになると考えられる．また，入力語に共通する関連エンティティが少ない場合も，ブートストラッピング法を用いて入力の数を増やすことで，入力（再入力を含む）に共通する関連エンティティの範囲を拡大できると考えられる．

図 3 は，Wikipedia Sets にブートストラッピング法を適用したときのモデルである．反復回数を N とし，I 回目の反復において，出力の上位 I エンティティを再入力させる．このように徐々に再入力のエンティティ数を増加させることで，関連エンティティの精度を維持したまま網羅性を向上できると考えられる．

ただし，このブートストラッピング法では，再入力するエンティティが入力として適切であるという仮定に基づいているため，再入力に用いるエンティティ数や反復回数などのパラメータに注意する必要がある．本研究の Wikipedia Sets では，上記の単純化したブートストラッピング法を用いるが，今後，各パラメータを適切に設定するための方法を検討する必要がある．

5.　評　　価

5.1　評 価 環 境

提案手法の有効性を検証するために，被験者を用いた評価を行った．関連度に関する評価方法は未だ一般的な方法が確立されていないため，文献 [19] と同様に，本研究でも，被験者に関連しているかどうかを判定してもらう方法を用いた．評価方法として，あらかじめ用意した複数の入力語に対して，Wikipedia Sets（ブートストラッピング法を適用，ブートストラッピング法なし）を用いて関連エンティティを取得し，被験者にどの程度関連しているかを判定してもらうという方法を採用した．なお，ブートストラッピング法を適用した場合の反復回数 N を 5 とした．比較手法として，連想検索サービスとして Web 上で公開されている Preferred Infrastructure 社の連想検索サービス reflexa の他，SEAL [14]，Google Sets を用いた．SEAL や

表 1　評価に用いた入力語句群と想定されるコンテキスト
Table 1　Input words and their contexts for evaluation

評価に用いた入力語句群	想定されるコンテキスト
リクナビ - マイナビ	就職活動に関するもの
リクナビ - 日雇い	アルバイトに関するもの
ナイフ - キャンプファイヤー	サバイバルに関するもの
ナイフ - 包丁	刃物に関するもの
ナイフ - トランプ - マッチ	手品に関するもの
アップル - 携帯音楽プレーヤー	iPod
アップル - 携帯電話	iPhone
アップル - UNIX	Mac OS X
大阪大学 - 名古屋大学	旧帝国大学や国立大学
大阪大学 - 追手門学院大学	大阪にある大学

Google Sets は類似した語句を抽出するシステムであるが，入力語が何らかのコンテキストを共有している場合，入力語とコンテキストが類似した語句の抽出，すなわち連想される語句の抽出も可能であると考えたため，比較手法として採用した．なお，reflexa や Google Sets の詳細なアルゴリズムは公開されていないが，reflexa は ESA [15] の関連度計算の手法（Wikipedia に出現する語を Wikipedia の記事でベクトル化する方法）を応用させた手法を用いていることは公開されている．

まず，評価に用いる入力語を決定する必要があるため，3 人の被験者から聴取を行った．具体的には，被験者に 2 種類以上のコンテキストを持つ名詞を挙げるよう指示し，それぞれコンテキストを構成する入力語句群と，入力から想定されるコンテキストについて聴取した．その結果，表 1 に示す 10 の入力語句群と想定されるコンテキストを得た．なお，想定されるコンテキストは被験者の回答によるものであるが，これに限定するものではない．次に，前述の被験者を 1 人含む被験者 5 人を用いて，各手法を用いて取得した関連エンティティを評価した．表 1 の 10 の入力語句群から，各手法を用いて連想されるエンティティ（あるいは語句）を抽出し，上位 30 件ずつ取り出した．このとき，被験者の先入観を無くすために，各手法によって取得した関連エンティティ（語句）を混在させ，ランダムに並び替えたリストを作成した．このリストを被験者に提示し，入力語句群に対して，それぞれの関連エンティティ（語句）がどの程度連想されるかを，文献 [19] の方法に従い，被験者が 0（全く連想されない）から 4（強く連想される）の 5 段階で判定した．注意事項として，被験者には入力語句群のコンテキストを考慮した上で判断するよう指示した．また，被験者の思考を狭める可能性があるため，想定されるコンテキストは被験者に提示しなかった．

5 人の被験者による評価を行った後，5 段階の判定の平均値を関連度として，各手法で取得した関連エンティティ（語句）の順位に対して $nDCG_p$ 値（ただし $p > 1$）を算出した．$nDCG_p$ 値は精度と網羅性の双方を考慮した評価指標であり，順位付きのリストを評価するためによく用いられる [20].

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{2}$$

ただし，

$$DCG_p = R_1 + \sum_{i=2}^{p} \frac{R_i}{\log_2 i} \tag{3}$$

$$IDCG_p = 4 + \sum_{i=2}^{p} \frac{4}{\log_2 i} \tag{4}$$

なお，$IDCG_p$ は，抽出した上位 p の関連エンティティ（語句）がすべて強く連想される (5 段階評価で最大値の 4) と判断された場合の DCG_p である．今回の評価では，$p = 10, 20, 30$ のときについて，それぞれ $nDCG_p$ を算出し，評価指標とした．

5.2 評価結果

$nDCG$ 値による評価結果を表 2 に示す．なお，提案手法の Wikipedia Sets において，ブートストラッピング法を用いた場合を With BS，ブートストラッピング法を用いなかった場合を Without BS と表記している．また，括弧付きの数字は規定の数 (p) の出力が得られなかった場合に，不足分を，全く連想できない（すなわち判定が 0 である）とみなして $nDCG_p$ を計算したときの値である．ハイフンは全く出力が得られなかった，もしくは出力数が増加しなかった場合である．

表 2 より，提案手法の Wikipedia Sets が多くの入力例に対して，他の手法よりも，コンテキストに沿って連想されるエンティティを抽出できていることがわかる．reflexa と比較すると，提案手法はほとんど全ての入力において同等かそれ以上の評価値を出しており，提案手法のほうが優れているといえる．reflexa と提案手法は共に Wikipedia をデータソースとして構築されたシステムであることから，提案手法のアルゴリズムとしての有効性を確認できる．また，SEAL や Google Sets は，入力語句が類似していることを発見できれば，そこから他の類似語句を抽出することによってコンテキストに依存した関連語を取得できる一方，入力語句の類似点を発見できなかった場合，全く出力が得られないという現象が発生した（表 2 のハイフン）．一方，提案手法は入力語句に類似点がなくても，入力語句から連想できる共通のエンティティがあれば出力が得られるように設計されているため，評価で用いた全ての入力語句に対してコンテキストに依存した関連エンティティを発見できている．

また，ブートストラッピング法を用いた場合のほうが，用いなかった場合と比較して評価値が高く，特に $p = 30$ の場合において大きな差が生じている．これは，入力語に共通する関連エンティティが少ない場合，ブートストラッピング法を用いない手法では，出力の下位の関連エンティティがコンテキストを考慮していないエンティティ（入力語のうちの一つとのみ関連しているエンティティ）となるのに対し，ブートストラッピング法を用いた手法では，出力の上位の関連エンティティをフィードバックさせることによって入力の数が増え，結果として入力語（再入力を含む）に共通する関連エンティティの数が増加したためであると考えられる．ただし，初期状態（反復を開始する前）の出力における（最）上位の関連エンティティが，コンテキストに沿って連想されるエンティティとして正しいという仮

表 2　評価結果 ($nDCG_p$)
Table 2　Evaluation results ($nDCG_p$)

TOP 10 ($p = 10$)

入力語句	Wikipedia Sets		reflexa	SEAL [14]	Google Sets
	With BS	Without BS	(ESA [15] based)		
リクナビ - マイナビ	**0.860**	0.795	0.745	0.574	0.769
リクナビ - 日雇い	0.580	**0.651**	0.607	(0.543)	0.529
ナイフ - キャンプファイヤー	**0.638**	0.453	0.322	0.574	0.262
ナイフ - 包丁	**0.713**	0.701	0.547	0.254	0.611
ナイフ - トランプ - マッチ	0.284	**0.292**	0.170	0.262	(0.197)
アップル - 携帯音楽プレーヤー	0.654	**0.666**	0.574	0.455	0.590
アップル - 携帯電話	0.456	0.500	0.477	0.626	**0.656**
アップル - UNIX	0.605	**0.625**	0.516	-	0.604
大阪大学 - 名古屋大学	0.748	0.766	0.581	**0.783**	0.725
大阪大学 - 追手門学院大学	0.646	0.599	0.617	0.410	**0.679**
平均	**0.618**	0.605	0.516	0.492	0.603

TOP 20 ($p = 20$)

入力語句	Wikipedia Sets		reflexa	SEAL [14]	Google Sets
	With BS	Without BS	(ESA [15] based)		
リクナビ - マイナビ	**0.776**	0.756	0.648	0.593	0.666
リクナビ - 日雇い	0.586	**0.617**	0.590	-	0.494
ナイフ - キャンプファイヤー	**0.587**	0.431	0.317	0.514	0.254
ナイフ - 包丁	**0.652**	0.629	0.461	0.221	(0.485)
ナイフ - トランプ - マッチ	0.249	0.243	0.152	**0.293**	-
アップル - 携帯音楽プレーヤー	0.641	**0.651**	0.547	0.410	0.508
アップル - 携帯電話	0.421	0.452	0.457	0.535	**0.573**
アップル - UNIX	**0.587**	0.585	0.476	-	0.560
大阪大学 - 名古屋大学	0.728	**0.742**	0.576	0.686	0.647
大阪大学 - 追手門学院大学	0.621	0.561	0.571	0.434	**0.678**
平均	**0.585**	0.567	0.479	0.461	0.548

TOP 30 ($p = 30$)

入力語句	Wikipedia Sets		reflexa	SEAL [14]	Google Sets
	With BS	Without BS	(ESA [15] based)		
リクナビ - マイナビ	**0.730**	0.656	(0.532)	0.487	0.657
リクナビ - 日雇い	0.571	**0.576**	0.537	-	0.482
ナイフ - キャンプファイヤー	**0.529**	0.380	0.273	0.471	0.225
ナイフ - 包丁	**0.626**	0.604	0.429	0.208	-
ナイフ - トランプ - マッチ	0.232	0.237	0.125	**0.296**	-
アップル - 携帯音楽プレーヤー	**0.598**	0.574	0.501	0.388	0.493
アップル - 携帯電話	0.416	0.410	0.442	0.463	**0.558**
アップル - UNIX	0.550	**0.557**	0.475	-	0.507
大阪大学 - 名古屋大学	0.692	**0.703**	0.561	0.635	0.608
大阪大学 - 追手門学院大学	0.569	0.547	0.569	0.415	**0.689**
平均	**0.551**	0.524	0.435	0.420	0.527

定に基づいている．そのため，初期状態ですでに上位にコンテキストと関連のないエンティティが含まれている場合は，反復処理によってコンテキストから離れた関連エンティティを取得してしまうことになる．ブートストラッピング法は，反復の停止条件，各反復において再入力に用いる関連エンティティの数など，パラメータによって性能が大きく影響を受けるため [21]，これらを自動的に調整するような仕組みが必要である．

5.3　出力の比較

より詳細に結果を分析するため，各手法によって得られた出力を比較した．「リクナビ」「マイナビ」と「リクナビ」「日雇い」をそれぞれ入力語句とした場合の各手法の出力の上位 15 件の比較を表 3 に示す．表 3 より，提案手法である Wikipedia Sets が，入力語の代表的なコンテキスト（表 1 より就職活動に関するものとアルバイトに関するもの）を考慮した上で関

表 3　出力の比較
Table 3　Comparison of outputs

入力語句: リクナビ - マイナビ				
Wikipedia Sets		reflexa	SEAL [14]	Google Sets
With BS	Without BS	(ESA [15] based)		
就職活動	就職活動	リクナビ	ブンナビ	マイナビ
合同企業説明会	合同企業説明会	マイナビ	学情ナビ	リクナビ
リクナビ	インターンシップ	毎日就職ナビ	就活ナビ	転職
転職	マイナビ	学情	日経就職ナビ	就職
インターンシップ	リクナビ	日経ナビ	就活ラボ	アルバイト
エントリーシート	求人広告	新卒	ネオキャリア就職ナビ	毎日就職ナビ
マイナビ	リクルート	転職	ベンチャー就職ナビ	派遣
履歴書	毎日コミュニケーションズ	毎日コミュニケーションズ	日経	求人
リクルート	転職	ナビ	外資系	仕事
適性検査	労働者派遣事業	就職	女の	紹介転職
企業	正社員	人材	エンジャパン	日経就職ナビ
面接	エントリーシート	キャリア	パッション就職ナビ	お役立ち
証明写真	面接	向	日経ナビ	求人情報
フリーター	適性検査	サイト	塾講師ナビ	独立
学情	圧迫面接	派遣	学情	スキルアップ

入力語句: リクナビ - 日雇い				
Wikipedia Sets		reflexa	SEAL [14]	Google Sets
With BS	Without BS	(ESA [15] based)		
就職活動	就職活動	日給	人材	リクナビ
リクナビ	リクルート	求人	紹介予定	日雇い
リクルート	非正規雇用	終身雇用	特定労働者	毎日就職ナビ
合同企業説明会	労働者派遣事業	求職	一般労働所	日経就職ナビ
労働者派遣事業	正社員	転職	登録型の日雇い	日経ナビ
履歴書	アルバイト	賃金	短期での	みんなの就職活動日記
インターンシップ	ブルーカラー	手当	登録型	学情就職 navi
転職	公共職業安定所	人件	二重	非正規雇用
エントリーシート	フリーター	雇用	インテリジェンスで製造	アクセス就職ナビ
フリーター	雇用	職種	-	転職
公共職業安定所	就職難	機密	-	学情ナビ
求人	株式会社	就業	-	リクルートナビ
非正規雇用	職業	就職	-	応募方法
適性検査	求人	労働	-	ダイヤモンド lead 就活ナビ
正規雇用	正規雇用	厚生	-	貧困ビジネス

連エンティティを精度良く取得できていることがわかる．例えば,「リクナビ」と「マイナビ」を入力語とした場合，就職活動（特に新卒採用）に関連のある「エントリーシート」や「インターンシップ」といったエンティティが抽出されているのに対し,「リクナビ」と「日雇い」を入力語とした場合，新卒採用とは異なる側面の就職活動として「非正規雇用」や「公共職業安定所」といったエンティティが抽出されている．比較手法でも，ある程度コンテキストを考慮した関連語を抽出できているが，SEAL や Google Sets，reflexa は並列関係にある語句を同時に取得しがちである．一方，提案手法では，並列関係以外にも様々な関係にある関連エンティティをバランスよく取得できていることがわかった．これは，既存の類似語句セットを抽出するような研究とは別の価値があることを意味している．

6. まとめと今後の課題

本研究では，複数の語句が入力として与えられた時に，その語句群から構成されるコンテキストを推測し，関連するエンティティ集合を抽出する手法を提案した．本手法では，複数の入力語に対し，それぞれの関連エンティティと関連度を取得した後，共通する関連エンティティの関連度をマージして出力する．また，得られた出力を再入力するというプロセスを反復すること（ブートストラッピング法）により，入力語が構成するコンテキストを強調させ，出力の精度及び網羅性向上を図る．評価により，提案手法が複数の入力語に対して，それら入力が構成するコンテキストを考慮して連想関係を取得できることを確認した．また，ブートストラッピング法の有効性を確認した．提案手法は，SEAL や Google Sets などの既存の類似語句セッ

トを抽出するような研究とは異なり，並列関係にあるエンティティ以外にも様々な種類の関連エンティティを抽出できるため，連想検索やクエリフリー検索，文書分類などのアプリケーションの基盤技術として利用できると考えられる．

今後の課題として，関連度のマージ方法やブートストラッピング法におけるパラメータの決定方法の検討が挙げられる．本研究では，関連度のマージ方法として，関連度を単純に加算する方法を用いたが，考えられる他の方法との比較を客観的に行い，最適なマージ方法を検討する予定である．また，ブートストラッピング法では，反復回数や再入力のエンティティ数を固定していたが，出力結果の変化の度合をみながら，これらのパラメータを自動で決定する仕組みを検討している．

また，出力の規模拡大（およびそれに伴う計算量増加への対応）が挙げられる．現状では，Wikipedia シソーラスを利用して，複数語の入力に対して共通する関連エンティティを取得している．しかし，Wikipedia シソーラスは複数語の入力を想定して構築されておらず，一つのエンティティに対して，強く関連しているエンティティのみ（平均 160 件程度）を定義しているため，共通して連想されるエンティティを取得するというアプローチでは，取得できる連想関係が限定されてしまう場合がある．この問題に対して，より多くの連想関係を定義できるように Wikipedia シソーラスを再構築する方法が考えられるが，保持すべき情報が膨大になる，すなわち，個々のエンティティに対して関連エンティティを取得するためにかかる計算量が膨大になるという問題がある．このような問題を解決するため，連想関係の裏にあるコンテキストがどのような種類であるかを，オントロジや Web などの外部情報を用いて事前に把握することを検討している．また，大規模な情報を高速に取り扱うアルゴリズムを模索する予定である．

謝辞 本研究の一部は，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) 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] P.A. Chirita, C.S. Firan, and W. Nejdl, “Personalized Query Expansion for the Web,” Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp.7–14, July 2007.

[2] R. Mandala, T. Tokunaga, and H. Tanaka, “Query Expansion using Heterogeneous Thesauri,” International Journal of Information Processing and Management, vol.36, no.3, pp.361–378, May 2000.

[3] R. Kraft, F. Maghoul, and C.C. Chang, “Y!Q: Contextual Search at the Point of Inspiration,” Proceedings of International Conference on Information and Knowledge Management (CIKM), pp.816–823, Oct./Nov. 2005.

[4] D. Shen, Z. Chen, Q. Yang, H.J. Zeng, B. Zhang, Y. Lu, and W.Y. Ma, “Web-page Classification through Summarization,” Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp.242–249, July 2004.

[5] H. Schütze, and J.O. Pedersen, “A Cooccurrence-based Thesaurus and Two Applications to Information Retrieval,” International Journal of Information Processing and Management, vol.33, no.3, pp.307–318, May 1997.

[6] Y. Jing, and W.B. Croft, “An Association Thesaurus for Information Retrieval,” Proceedings of Recherche d’Information Assistée par Ordinateur Conference (RIAO), pp.146–160, Oct. 1994.

[7] 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.

[8] 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.

[9] H. Chen, T. Yim, D. Fye, and B. Schatz, “Automatic Thesaurus Generation for an Electronic Community System,” Journal of the American Society for Information Science, vol.46, no.3, pp.175–193, Apr. 1995.

[10] C.J. Crouch, “A Cluster-based Approach to Thesaurus Construction,” Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp.309–320, June 1988.

[11] Z. Chen, S. Liu, L. Wenyin, G. Pu, and W.Y. Ma, “Building a Web Thesaurus from Web Link Structure,” Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp.48–55, July/Aug. 2003.

[12] A. Budanitsky, and G. Hirst, “Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures,” Proceedings of Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), June 2001.

[13] Z. Ghahramani, and K.A. Heller, “Bayesian Sets,” Proceedings of Advances in Neural Information Processing Systems (NIPS), Dec. 2005.

[14] R.C. Wang, and W.W. Cohen, “Language-Independent Set Expansion of Named Entities using the Web,” Proceedings of International Conference on Data Mining (ICDM), pp.342–350, Oct. 2007.

[15] E. Gabrilovich, and S. Markovitch, “Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis,” Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp.1606–1611, Jan. 2007.

[16] T. Yamamoto, S. Nakamura, and K. Tanaka, “TermCloud for Enhancing Web Search,” Proceedings of International Conference on Web Information Systems Engineering (WISE), pp.159–166, Oct. 2009.

[17] 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.

[18] 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.

[19] J. Gracia, and E. Mena, “Web-based Measure of Semantic Relatedness,” Proceedings of International Conference on Web Information Systems Engineering (WISE), pp.136–150, Sept. 2008.

[20] K. Järvelin, and J. Kekäläinen, “Cumulated Gain-based Evaluation of IR Techniques,” ACM Transactions on Information Systems, vol.20, no.4, pp.422–446, Oct. 2002.

[21] 小町守，工藤拓，新保仁，松本裕治，“Espresso 型ブートストラッピング法における意味ドリフトのグラフ理論に基づく分析: 語義曖昧性解消における評価,” 人工知能学会論文誌，vol.25, no.2，pp.233–242，Jan. 2010.