知能と情報（日本知能情報ファジィ学会誌）Vol.22, No.6, pp.707－719（2010）

原著論文

WWW活用による語の比喩的素描手法†

桝井 文人*1・ジェプカ ラファウ*2・木村 泰知*3・福本 淳一*4・荒木 建治*2

論文では，クエリ語に対して説明文や定義文を回答する代わりに，WWWから収集した断片知識を使って比喩的に素描する手法を提案する．提案手法は，直喩表現を生成する指標パターンを利用して，WWWから大量の名詞句の関係を収集する．そしてこれらの情報に基づいて，デスクリプタと呼ばれる，クエリ語を描写する断片知識を獲得する．各デスクリプタの一般性と局所性を考慮してその記述力を推定し，これに基づいてランキングして視覚化する．さらに，比喩的関係が持つ性質と複数の定型パターンによる知識獲得技術を組み合わせることによって，獲得したデスクリプタ集合を上位語，属性語に分類する．ユーザは視覚化された比喩的素描を見ることによって，連想的にクエリ語の意味を把握できる．

提案手法の有効性を検証するために，実装システムMurasakiを構築し，いくつかの評価実験を実施した．その結果，基本性能については，bag of wordsを用いるよりもかなり有効であることを確認した．また，検索サイトにおける注目キーワードに対する応答性能では，汎用辞書の性能を大きく上回り(60%の適合率)，新語や固有名詞に対して非常に有効であることがわかった他，ランキング性能(74%のMRR)や分類性能(63%の成功率)についても有効性が確認できた．さらに，獲得するデスクリプタの網羅性を安定させれば，比喩的素描によってWikipediaに近い効果が得られる可能性(81%のヒット率)を持つこともわかった．この結果は，我々が提案する比喩的素描による表現手法が，説明文や定義文の代わりに機能する可能性を示したとともに，既存の汎用辞書には対応できない新語や固有名詞に対しても有効であることを示している．

キーワード：比喩的素描，デスクリプタ，直喩表現，定型パターン

1．はじめに

本論文では，クエリ語に対して説明文や定義文を回答する代わりに，WWWから収集した断片知識を使って比喩的に素描する手法を提案する．

ユーザの情報要求に回答する代表的な技術として情報検索が普及している．情報検索では，ユーザの要求に対するレスポンスはあくまでも大雑把に適合する文書の提示にとどまる．これに対して，ユーザの要求に関して答えそのものを返す技術として，質問応答が注目されている．

質問応答では答えそのものを回答する．最も基礎的なものとして，答えの範囲が名詞句や固有表現に限定されるfactoid型タスク[1，2，3，4]がある．さらに，より実用的かつ高次な質問へ対応するために語や事物の意味を問うdefinition型[3]，理由を問うwhy型[5]，列挙された一連の質問に回答する対話指向型(IAD)[6]などのタスクについての研究が進められている．

このような高次タスクに対しては，factoid型質問応答技術の延長上の技術論では有効な結果は得られないことが指摘されている[7]．これらに対応するためには，深い解析処理や文脈処理，自動要約技術[8]などが必要となる．

しかしながら，現時点では，必要とされる個々の技術自体が実現難易度の高い研究課題であるため，これらの高度技術を組み合わせることによって高い処理精度を確保できる可能性は低い．さらに，上記の複雑な処理の組み合わせは計算量増加の要因となり，動的な処理を考えた場合にも不利となる．

また，ネットワークを介して大量のデータが瞬時に世界を駆け巡る今日においては，適切な状況判断のために保持する情報の鮮度を保つ必要もある．そのためには，最近まで存在しなかった新語や注目を集めている語に対する知識を抽出整理できる即応性と網羅性を兼ね備えた知識獲得機構が必要となる．

そこで本論文では，概念に対する定義を過不足なく

† WWW-based Figurative Descriptions for Japanese Word
Fumito MASUI, Rafal RZEPKA, Yasutomo KIMURA, Jun-ichi FUKUMOTO and Kenji ARAKI

*1 北見工業大学 工学部 情報システム工学科
Department of Computer Science, Kitami Institute of Technology
*2 北海道大学大学院 情報科学研究科 メディアネットワーク専攻
Graduate School of Information Science and Technology, Hokkaido University
*3 小樽商科大学 商学部 社会情報学科
Department of Information and Management Science, Otaru University of Commerce
*4 立命館大学 情報理工学部 メディア情報学科
Department of Media Technology, Ritsumeikan University

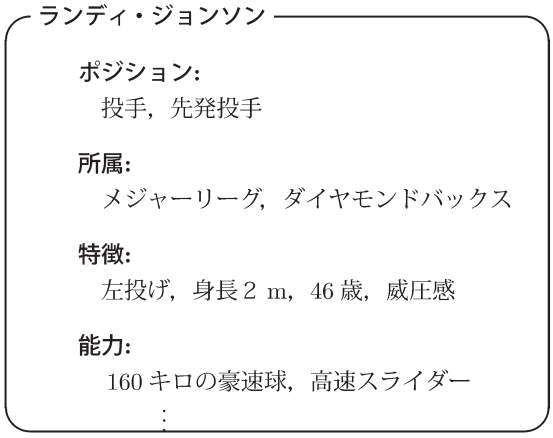

図１　理想的な比喩的素描の例

回答するのではなく，クエリの定義がイメージでき，理解を促すに十分な補足知識を提供でき，新語や注目されいている語についても追従する手法の構築を目指す．例えば，「ランディ・ジョンソン」という語は一般的な辞書には登録されていないが，これを素描する断片知識として，図１に示すようなものが挙げられる．これらの断片知識集合を提示することによって，ユーザに連想に基づく理解を促し，結果的に定義文を示す場合と同等の効果が期待できる．

提案手法は，桝井らによる比喩性検出のための知識洗練手法[9]を拡張し，WWWより断片知識の獲得と洗練を行なう．比喩検出のメカニズムを応用することにより比較的単純な処理でWWWから対象の総体を素描する断片知識を効率よく獲得・整理できる．そしてさらに，複数の定型パターンと適合フィードバックの組み合わせにより，獲得した知識分類を行なう．

Chinchor[10]は，固有表現を叙述あるいは描写する知識を*descriptor*と呼んでいる．本論文で扱う断片知識も，対象を素描する知識である．よって，これらについてもデスクリプタと呼ぶことにする．

以下，２章で直喩関係の役割とデスクリプタの扱いを中心とした基本的なアイディアについて述べる．３章では，提案手法の詳細について説明し，４章，５章でいくつかの評価実験とその結果に対する考察を行なう．

２．直喩関係の役割とデスクリプタの扱い

本章では，比喩的関係が情報要求への応答において取り得る役割について議論する．

比喩とは，あることがらをそれと何らかの関係があるもうひとつのことがらに例えて表現する手法である[11，12]．言語形式に基づくと，比喩表現は，喩辞・被喩辞・指標の三要素に分解でき，喩辞と被喩辞の意味的な比較によって解釈される[13]．これらの比喩的関係は新聞記事などの実用文やWeb文書中にも多く出現し[11]，普通名詞だけでなく固有名詞が対象要素となることも多い．

その中でも，直喩表現はすべての要素が明示され，「ランディ・ジョンソンのような投手」などのように形式化される場合が多い．このような形式化された表現において比喩指標は定型パターンとして扱うことができる．

「*A*のような*B*」というパターンを用いた直喩表現の例を表１に示す．これらはいずれも比喩あるいは例示であり，喩辞と被喩辞を比較することで解釈される．注目すべきことに，これらの表現を理解する際には，比較される要素が比喩的関係を持つというだけではなく，要素間には特定の関係が存在することも同時に解釈される．第一の例では，「ランディ・ジョンソン」と「投手」は上位−下位関係を持ち，第二，第三の例では，「ランディ・ジョンソン」と「威圧感」，「豪速球」は主体−属性の関係を持つ．

これらの表現を総合的に解釈すれば，「ランディ・ジョンソンは威圧感があり，剛速球を投げる投手である」といったイメージが掴める．そして，このような関係を特定するためのマーカとして，直喩表現を構成する比喩指標パターンを利用することは大いに有効である．

定型パターンを利用した知識獲得は近年注目されている手法である[14，15，16，17，18]．基礎的なものとしては，“A such as B”という定型パターンを用いてコーパスから収集した共起単語対の性能を確認したHearstの研究[14，15]や，複数の定型パターンを利用して得られる日本語名詞の上位下位関係の性能を調査した安藤らの研究[18]がある．大規模な試みとしては，複数の定型パターンを用いてWWWから大量の共起知識を収集してオントロジーを構築しようとするPantelらの研究[19]や，Shinzatoらの研究[20]などがあり，それぞれ従来の統計的手法より有効であると報告している．

上で示した研究は大規模な知識を取り出す点に主眼を置いているが，本研究は知識を取り出すだけではなく，それらを整理して素描しようとする点で目的が異なる．また，従来研究では様々な知識を獲得するために相当数のパターンを利用しているが，本研究では比

表１　比喩指標パターンを用いた直喩表現の例

言語形式	例	タイプ
喩辞＋指標＋被喩辞	ランディ・ジョンソンのような投手	上位語
喩辞＋指標＋被喩辞	ランディ・ジョンソンのような威圧感	属性（特徴）
喩辞＋指標＋被喩辞	ランディ・ジョンソンのような豪速球	属性（能力）

喩的関係の特性を利用することによって単一パターンを用いて複数種類の知識を取り出す．従来手法と比較してより単純なアプローチという点で効率的である．

桝井ら[9]は，定型パターンを利用した知識獲得技術と比喩指標の特性を組み合わせ，WWWからの属性値知識の抽出手法を提案している．

彼らの手法では，知識構築処理と知識洗練処理が独立して実施される．知識構築処理では，新聞記事コーパスに記述された形容詞−名詞の連体修飾関係を利用して語の共起(x_i, y_j)を抽出する．例えば，「赤いリンゴ」，「大きいリンゴ」からは(リンゴ，赤い)や(リンゴ，大きい)という共起が得られる．コーパス全体から収集した共起は名詞概念に関する属性値集合$x_i = \{y_1, y_2, \ldots, y_j, \ldots\}$として統合される．上の例では，リンゴ＝{赤い，甘酸っぱい，大きい，…y_j，…}となる．

知識洗練処理では，指標パターン「XのようにY」を利用する．(x_i, y_j)を指標パターンに適用し，「x_iのようにy_j」という比喩表現を生成する．これにより，「リンゴのように赤い」「リンゴのように甘酸っぱい」といった比喩表現が生成される．生成した表現のWWWにおける出現状況を参照すると適合性が判定できる．判定により，(リンゴ，大きい)は不適合となり属性値集合から削除される．以上の処理を全ての属性値に対して実施することで属性値集合が洗練される．

本研究では，対象とする知識はクエリ語とデスクリプタであるので，名詞−名詞の関係を扱う．上記の研究ではWWWから直接知識を抽出することが理論的に困難であると述べているが，我々の手法では，ブートストラップ的アプローチによってWWWからの直接的知識獲得を実現する．さらに，本手法で試みる獲得知識の分類も従来手法にはない点である．提案手法の実現にあたり，桝井ら[9]による指標パターン利用の考え方とスコアリングモデルを応用する．

以下，提案しようとする手法の基本的考え方を述べる．本手法では，指標パターンとして「XのようなY」を用いる．

Xにクエリ語q_iを適用して表現「q_iのような」を生成する．この表現を補完して「q_iのようなγ_j」を完成する語γ_jを収集するとクエリ語を表現するデスクリプタ候補集合$\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_j, \ldots\}$が得られる．

例えば，クエリ語が「ランディ・ジョンソン」の場合，「ランディ・ジョンソンのような」を生成し，WWWから「ランディ・ジョンソンのようなγ_j」をみたす語γ_jをデスクリプタ候補として収集すると，Γ＝{投手，豪速球，…}が得られる．

収集したΓの各要素を用いて「のようなγ_j」を生成する．この表現を補完して「$s_{j,k}$のようなγ_j」を完成する語$s_{j,k}$を調べ，異なり数kを得る．

上の例では，「のような投手」や「のような豪速球」を生成し，WWWから「$s_{投手,k}$のような投手」，「$s_{豪速球,k}$のような豪速球」を満たす語$s_{j,k}$の異なり数kを取得する．

上で述べたΓに関して得られる候補γ_jの頻度$freq(\gamma_j | x_i)$はクエリ語に対するデスクリプタの典型性の強さを示し，$q_{j,k}$に関して得られる異なり数kはクエリ語に対するデスクリプタの局所性の強さを示す．

上述の典型性と局所性に加えて，全ての「x_iのようなγ_j」の頻度がわかれば，文献[9]のスコアリングモデルに適用し，x_iに対するデスクリプタγ_jの記述力を推定し定量化できる．

次に，デスクリプタの分類について述べる．表現「リンゴのような果実」では，(リンゴ，果実)は上位−下位($is-a$)関係，表現「リンゴのような色」では，(リンゴ，色)は主体−特徴の関係，表現「リンゴのような頬」では，(リンゴ，頬)は連想関係を持つ．

これらの関係は，関係を明示的に示す他の定型パターンでも表現できる．例えば，上位−下位関係であれば，「x_iというγ_j」が生成できる．「リンゴという果実」のように表現できる．主体-特徴関係であれば，「x_iのγ_j」が生成でき，また，「リンゴのような頬」のように比喩のみで成り立つ関係は上記表現は生成できないため，他と区別できる．

以上より，クエリ語と各デスクリプタを，複数の定型パターンに適用して明示的な表現を生成し，それらの出現分布を調べれば，上記の関係を区別でき，その結果としてデスクリプタ集合の分類が可能となる．

3．提案手法

本章では，我々が提案する比喩的素描手法について，実装システムMurasakiの動作画面とともに詳述する．

提案手法の構成を図2に示す．本手法は，後述する9つの処理過程，(STEP1)クエリの取得，(STEP2)比較表現の生成(A)，(STEP3)表現検索，(STEP4)デスクリプタ候補の抽出，(STEP5)比較表現の生成(B)，(STEP6)表現検索，(STEP7)スコアの計算，(STEP8)デスクリプタ集合の構築，(STEP9)デスクリプタの分類，および(STEP10)整理と可視化から成る．

(STEP1)では，入力されたクエリx_iを認識する．このとき，x_iは形態素解析によって形態素の系列として認識される．

(STEP2)では，まず，指標パターンA「Xのような」を用意しておく．ここで，「XのようなY」は(X，ω，

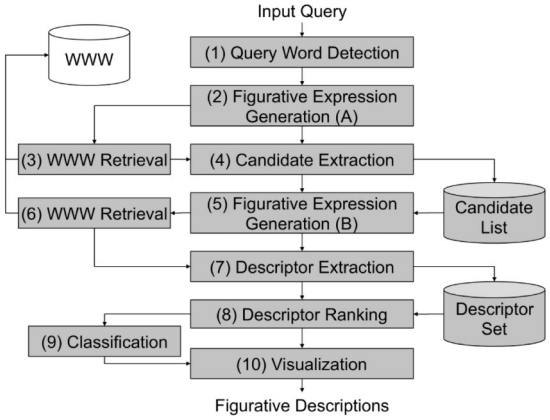

図 2　提案手法の構成概要

$Y|\omega$＝“のような”)のように表すことにする．パターンA「Xのような」を用いてパターンAのXにはクエリ語x_iを適用し，不完全な比較表現(x_i，ω，＊$|\omega$＝“のような”)，例えば「ランディ・ジョンソンのような」を生成する．

(STEP3)では，生成した表現(x_i，ω，＊$|\omega$＝“のような”)」をクエリとしてWWW検索を行なう．

(STEP4)では，検索された上位nページ中のsnippetを形態素解析し，{(x_i，ω，γ_1)，(x_i，ω，γ_2)，…，(x_i，ω，γ_j)，…}を検出する．さらにその構成要素となっている名詞句γ_jを抽出する．γ_jは「名詞」，「名詞連続」，「名詞＋の＋名詞」とする．

ここで解析対象をsnippetとしている理由は以下の通りである．本処理過程においては，デスクリプタ候補リストを作成することが主眼であるため，検索された比喩的表現のみが取得できればどのような文書に含まれているかは問題とはならない．したがって，検索クエリを含むsnippetを処理すれば目的は達成される．また，検索結果1,000ページ分のsnippetを対象とすれば，1,000以上の候補集合が得られることになり，網羅性は十分確保出来ると考えてよい．よって，検索結果1,000ページ分のsnippet集合から得られないデスクリプタ候補は，処理対象となっている比喩的表現として表現される可能性が極めて低いといえ，無視できる存在である．

なお，解析対象とする上位nページはデスクリプタ候補の規模を決定する値であり，規模を大きくすると素描の網羅性は向上するが処理時間が増大する．もちろんデフォルト値は設定されるが，網羅性と処理時間のバランスはユーザの意図や目的によって異なると思われるため，nはスライドバー操作によってユーザによる変更を可能としている．また，一般ユーザを想定した場合，「精度重視」や「網羅性重視」といった選択肢を提示するなど，より直感的インタフェイスを設定することも可能である．

以上より，デスクリプタ候補リストΓ＝{γ_1，γ_2，…，γ_j，…，γ_n}が得られる．「ランディ・ジョンソン」の例であればΓ＝{投手，威圧感，剛速球，…}となる．このとき，(x_i，のような，＊)の検索ヒット数も取得しておく．ここでいう検索ヒット数とは，検索結果として得られるページ数である[1]．したがって，同一ページ内に同じデスクリプタが複数出現した場合は1としてカウントする．

(STEP5)では，まず，「のようなΓ」という統語上のパターンBを用意しておく．パターンBのΓにはデスクリプタ候補γ_jを適用し，(＊，ω，$\gamma_j|\omega$＝“のような”)，例えば「のような投手」や「のような威圧感」といった不完全な比較表現を生成する．

(STEP6)では，生成した表現(＊，ω，$\gamma_j|\omega$＝“のような”)を検索クエリとしてWWW検索を行なう．

(STEP7)では，得られた検索ヒット数を生成表現の頻度$cf(\gamma_j)$として取得する．

同時に(x_i，ω，$\gamma_j|\omega$＝“のような”)のヒット数をx_iに対するγ_jの頻度$pf(\gamma_j|x_i)$として取得する．この値が許容範囲($pf(\gamma_j|x_i)\geq\alpha$)であった場合，$\gamma_j$はデスクリプタであると判断される．

以上の操作を全てのデスクリプタ候補に適用する．

(STEP8)では，クエリx_iを表現するデスクリプタ集合を構築する．取得した各デスクリプタについて，その取得頻度に基づいたスコアを計算する．この結果，候補リストΓの各要素γ_jについて，よりx_iと関連の強いものが把握でき，デスクリプタとしての妥当性を判定することができる．

以上の計算を全ての候補に適用し，クエリ語q_iに対するデスクリプタ集合を得る．

(STEP9)では，統語上のパターンを用いた三段階の処理によって，収集したデスクリプタを，「上位語」，「属性語」，「連想語(判定不可)」の三カテゴリに分類する(図3)．

まず第一に，パターン「XというY」に対してクエリ語q_iをXに，デスクリプタγ_jをYに適用して，(q_i，ω，$\gamma_j|\omega$＝“という”)，具体的には「ランディ・ジョンソンという投手」「ランディ・ジョンソンという威圧感」といった表現を生成する．生成した表現をクエリとしてWWW検索を行い，検索されたヒット数h_1を取

1　厳密にいえば「検索クエリ数」は「ページ数」と一致しない可能性が高いが，クエリ検出のためにページ解析を行なう際の計算量増大の影響は近似によって生じる誤差の影響よりも遥かに大きいためページ数を検索クエリ数として近似する．

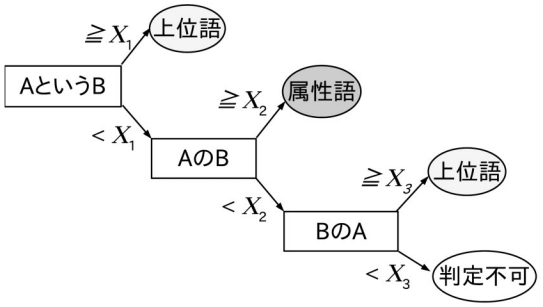

図3　分類処理の概要

図4　デスクリプタの出力結果

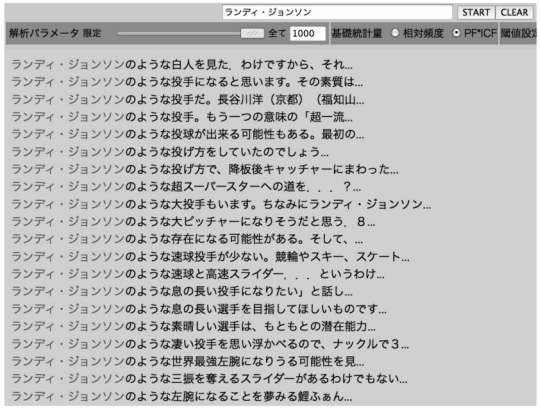

図5　抽出表現を含むパッセージ一覧

図6　デスクリプタ分類結果（上位語）

得する．ここで，h_1が条件($h_1 \leq X_1$)を満たせばγ_jを上位語と判定し，そうでない場合は次の判定処理を行なう．上の例では「投手」は上位語に分類される．

第二に，パターン「XのY」に対して，クエリ語q_iをXにγ_jをYに適用して(q_i，ω，γ_j|ω＝“の”)，例えば「ランディ・ジョンソンの剛速球」や「ランディ・ジョンソンのスーパースター」といった表現を生成する．前述の処理と同じ様にWWWより生成表現のヒット数h_2を取得する．h_2が条件($h_2 \leq X_2$)を満たせばγ_jを属性語と判定し，そうでない場合は次の判定処理を行なう．例では，「剛速球」が属性語に分類される．

第三に，前述のパターンに対して，クエリ語q_iをYにγ_jをXに適用して(γ_j，ω，q_i|ω＝“の”)，例えば「スーパースターのランディ・ジョンソン」や「肩甲骨のランディ・ジョンソン」といった表現を生成する．前述の処理と同じ様にWWWより生成表現のヒット数h_3を取得し，h_3が条件($h_3 \leq X_3$)を満たせばγ_jを上位語と判定する．そうでない場合は判定失敗と判断し，連想語として扱う．例では，「スーパースター」が上位語に分類される．

なお，各閾値は予備調査の結果から，X_1＝X_2＝X_3＝10を用いる．予備調査では，10種類のクエリ語とそれらの上位語，属性語合計101個について，三種類の表現パターン「AというB」「BのA」「AのB」として現れる出現分布を調べ，主観評価によって閾値を決定した．

(STEP10)では，デスクリプタ集合をスコアに基づいて整列させ，可視化表示する(図4)．提示されたデスクリプタにはリンクが付与されている．このリンクを辿ることによって，クエリ語とデスクリプタのAND検索結果が参照できる．

また，ユーザのメニュー選択によって，抽出表現を含むパッセージ一覧(パッセージビュー，図5)やデスクリプタ集合のリスト(リストビュー)，分類結果(カテゴリビュー，図6)などを表示する．グラフビューは素描されたデスクリプタ集合の全体像の把握，デスクリプタの影響力の強さを知りたい場合に有効であり，パッセージビューは各デスクリプタがどのような文脈中に出現しているかを知りたい場合に有効である．リストビューはどのようなディスクリプタが得られているか雑駁に把握したい場合に，カテゴリビューはデスクリプタの機能や特定の機能のみを把握したい場合に有効である．

・デスクリプタ集合構築のためのスコアの計算

デスクリプタ集合の構築における，各デスクリプタのスコア計算について説明する．

x_iを表現するデスクリプタ集合中のγ_jのスコア

$score(\gamma_j|x_i)$は，pf*icf[9]を用いて求める．x_iを素描するデスクリプタ集合におけるγ_jの典型性(強度)を示す$pf(\gamma_j|x_i)$と，x_i以外の語を素描する場合と比較してγ_jの局所性を示す$icf(\gamma_j)$の積として表す(式(1))．

$$score(\gamma_j|x_i) = \quad pf(\gamma_j|x_i) \ * \ icf(\gamma_j) \tag{1}$$

$icf(w_i)$は，WWW中に存在する全ての概念の総数Nと，w_iがデスクリプタと成り得る概念の数$cf(w_i)$の関係を正規化した値で，式(2)により求める．ただし，ここで$icf(w_i)$の真の値を得ることは計算コストから考えて非常に困難である．そのため，「のような」をWWW検索して得られる検索結果(*，ω，*|ω＝“のような”)のヒット件数をNとして用い，表現「のようなγ_j」をWWW検索して得られる検索結果(*，ω，$\gamma_j|\omega$＝“のような”)のヒット件数を$cf(\gamma_j)$として用いる．

$$icf(\gamma_j) = \quad log\frac{N}{cf(\gamma_j)} \ + \ 1 \tag{2}$$

クエリ語を「ランディ・ジョンソン」とした場合，処理過程(4)においてN＝85,300が得られ，処理過程(6)，(7)において$pf(\gamma_j|x_i)$および$cf(\gamma_j)$として次のような値が得られる．

γ_j	$pf(\gamma_j\|x_i)$	$cf(\gamma_j)$
威圧感	55	1860
肩甲骨	3	3
⋮	⋮	⋮

これらを式(1)と(2)に適用すると次にようにスコアが計算される．

$$score(\text{威圧感} \mid \text{ランディ・ジョンソン}) = \quad 55 \ * \ (log\frac{853000}{1860} + 1) = \quad 201.30 \tag{3}$$

$$score(\text{肩甲骨} \mid \text{ランディ・ジョンソン}) = \quad 3 \ * \ (log\frac{853000}{3} + 1) = \quad 19.35 \tag{4}$$

さらに，STEP10において整列した比喩的素描を可視化する際には，デスクリプタのスコア合計が1となるように各スコアを正規化する．

4．評価

提案手法の有効性を検証するために，Murasakiとベースラインシステムとの比較評価，ならびにMurasakiが出力したデスクリプタ集合のランキング性能，および分類性能の分析評価を行なった．以下，各評価環境および評価結果とその考察について詳述する．

・適切性に対する評価

獲得したデスクリプタの精度を検証するために，人手による評価を行なった．評価は被験者20名を用い，被験者には自由に考えたクエリ語をMurasakiに入力し，クエリ語に対して提示されたデスクリプタ集合それぞれの正否判定をしてもらった．被験者は，工学系学部4年次生および大学院博士前期課程学生で，今回の実験以外には本研究とは関わりを持たない学生である．なお，前章(4)の対象ページ数はn＝1,000とした．これは，システムが用いるWeb検索エンジン(Yahoo!API)が検索結果として出力するページ数の上限が1,000ページとなっており，その最大数を採用したためである．

判定において，

(正)クエリ語について描写されている．
(否)クエリ語について描写されていない．

という基準で，できるだけ客観性を保持して二値判断をしてもらった．その結果，173個のクエリ語について判定結果が得られた．

比較のために，単純な共起情報に基づいてデスクリプタ相当の知識を提示するベースラインシステムを実装し，同様の評価を実施した．ベースラインでは，クエリ語x_iを先端記号とするサイズm形態素分の窓を設定し，窓内に現れる最も近傍の名詞句ϵが存在する場合に，ϵをx_iのデスクリプタ相当の共起要素として抽出し，頻度情報を取得する．得られた頻度情報を用いて，提案手法と同じスコア計算方法を適用し，ランキングを行い，結果を出力する．

なお，ウィンドウサイズについては，m＝2～12として各出力に対して評価結果を得た．

次に，ベースラインに対する提案手法の相対再現率[21]を求め，提案手法の網羅性を検証した．相対再現率とは，本来ユーザが問題解決に必要とする正解数の何割をシステムが提示できたかを表す尺度であり，再現率計算が困難な場合に正解データの網羅性を吟味するために有効な評価尺度である[22]．

・即応性に対する評価

提案手法の即応性を検証するために，一般的なキーワードを用いて評価実験を実施した．まず，以下に示す主要なWWW検索サービスにおける注目キーワードとして，(a)「Yahoo!JAPAN急上昇キーワード」全30語，(b)「はてな注目キーワード」全20語，(c)「gooのウェブ検索急上昇キーワード」全10語，(d)「Google急上昇キーワード」全10語(2010年5月10日付)を取得した．サイト毎に取得したキーワードの語数が異なっているが，これは各サイトトップページにて公開されて

いるキーワード全てを対象としたためである．

なお，(a)には複数の単語から構成されるキーワードが存在していた．複数キーワードへの処理については，以下のような問題が生じる．まず，キーワード個別の比喩的関係を扱うべきなのか，与えられたキーワードセットをひとつの意味表現として比喩的関係を考えるべきなのかという問題を考える必要がある．さらに，個別の素描を単純に連結すればよいのか，あるいは各素描を融合する必要があるのか，についても議論する必要がある．

以上のことから，複数キーワードからなるクエリ9語の処理は対象外として除外し，合計61語(異なり数60語)を実験対象とした．

各キーワードをMurasakiに適用し，得られたデスクリプタを評価した(対象ページ数$n=500$)[2]．また，比較対象として汎用辞書[3]とWikipediaを用い，上記キーワードがエントリされているか否かを調べた．

・ランキング性能に対する評価

次に，デスクリプタ集合のランキング性能を評価した．まず，前節で挙げた173クエリより70個を無作為に抽出した．それらに対して提案手法が提示したデスクリプタ上位5位までの集合を対象として，人手によって妥当と判定された最高順位のデスクリプタを対象としてMRRを計算した．

MRRは，順位付けされたシステム出力において最高位の正解順位の逆数(RR：Reciprocal Rank)の平均値として表現される．入力クエリについて提示されたデスクリプタをスコアに基づいて順位付けして得たMRR値は，各クエリを表現するデスクリプタ集合の平均的な抽出性能を評価する指標であるとみなせる．この操作を50回繰り返し，MRRの平均値を求めた．

・分類性能に対する評価

デスクリプタの分類性能を評価した．まず，上述した173クエリより57語(普通名詞28語，固有名詞29語)を選択した．普通名詞と固有名詞をそれぞれ30語程度取り出した．30語程度とした理由は，統計学的考察を行なう上で最低限必要とされるサンプル数(magic number)を確保したいという意図による．多義性解消に由来する問題を回避するため，多義性を持つ語を除外した結果，最終的に57語となった．

選んだクエリをMurasakiに適用してデスクリプタの獲得と分類処理を行なった(対象ページ数$n=1{,}000$)．判定では，EDR概念辞書[23]を用い，辞書にエントリがあるものについてはその意味素性とその関係を利用して判定した．判定にあたり，デスクリプタと概念辞書エントリの文字列が完全に一致しない場合であっても同義であると判断できる場合はエントリされた語として扱った．辞書にエントリがないものについては，クエリとデスクリプタを用いて人手にてWWW検索を行い，検索された文書内のクエリを説明する文脈内でデスクリプタが使われているかどうかを確認して判断した．判定は「適切(十分適切○，適切△)」，「不適切(×)」とした．

4.1 評価結果

適切性に関する評価結果のグラフを図7に示す．グラフは，デスクリプタ抽出性能に関して，ベースラインと提案手法の適合率を比較したものである．

ベースライン結果の適合率は，0.3〜0.4(適合率平均は0.35)であったのに対し，提案手法の適合率は0.53となり，ベースラインの結果を全て上回った．

適合率が最も高いウィンドウサイズ$m=6$のベースライン結果を対象として提案手法の相対再現率1を各クエリに対して計算したところ，相対再現率は1.11と

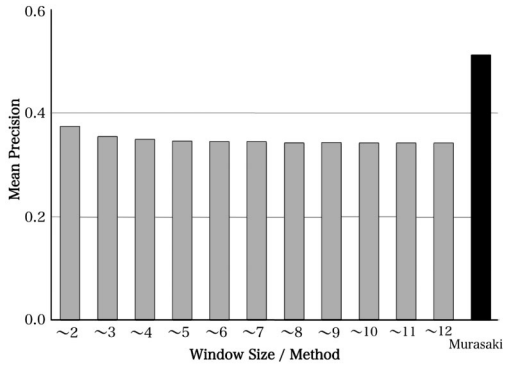

図7 平均適合率の比較

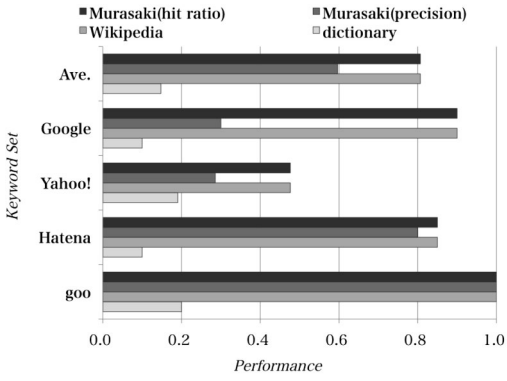

図8 注目キーワードに対する出力の即応性の比較

2 $n=100$〜1,000とした予備実験により，500ページを超えると取得できるデスクリプタ数が急激に鈍化し，新たなデスクリプタがほとんど得られなくなることがわかったため，nの最適値を500とした．

3 デジタル大辞泉(23万語収録)，小学館

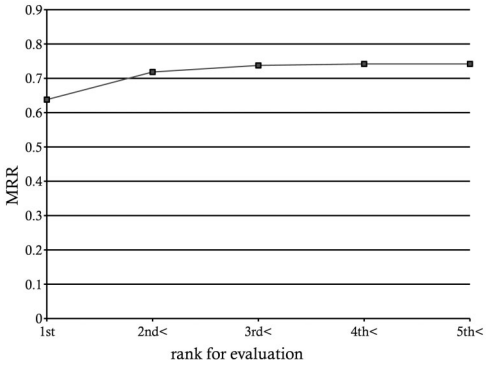

図 9　評価対象の範囲と平均MRRの関係

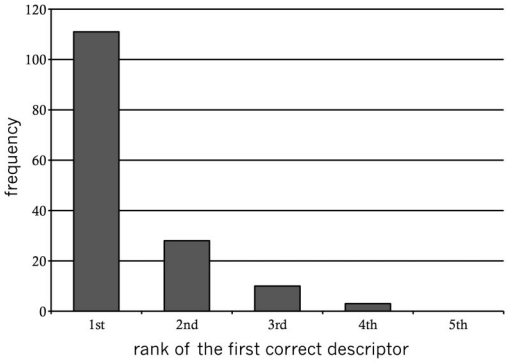

図10　妥当なデスクリプタの最高順位

なった．ベースラインは典型的なbag-of-words型アプローチであるため再現率が高くなる傾向が強いはずであるが，提案手法の再現率がベースラインを上回る結果となった．

即応性の評価結果を図 8 に示す．グラフでは，注目キーワードリスト(a)～(d)の評価結果とそれらの平均値を示している．「Murasaki(hit ratio)」は，デスクリプタが獲得できたクエリの割合，「Murasaki(precision)」は，デスクリプタを適切に獲得できたクエリの割合であり，「Wikipedia」はWikipediaにページエントリが存在したクエリの割合，「dictionary」は汎用辞書にエントリが存在したクエリの割合である．

平均値では，提案手法は，precisionでは0.60であったが，hit ratioでみると0.81とWikipediaと同じ数値を示した．汎用辞書では0.15であった．

キーワードリスト別にみると，(d)gooに対する結果が最も良い結果となり，Wikipediaと提案手法が全てに対応できた．逆に，(a)Yahoo!に対する結果が最も悪い結果となり，Wikipediaでは0.48，提案手法では0.29，汎用辞書では0.19であった．

次に，ランキング性能について述べる．図 9 は，上位 l 位を対象としたMRR(l =1，2，3，4，5)をグラフ化したものである．50回の試行に対するMRR平均値は0.74であり，MRR値は l = 3 付近でほぼ頭打ちとなっていることから，平均的には上位 3 位までに妥当なデスクリプタを提示できたといえる．この傾向は，妥当なデスクリプタが得られた最高順位の分布(図10)からも確認できる．

分類性能の評価結果を表 2 に示す．全体の平均適合率は0.63であった．上位語への分類性能は0.56，属性語への分類性能は0.69であり，属性語への分類性能が上位語を上回った．普通名詞と固有名詞の区別でみると，普通名詞では0.51，固有名詞では0.74と，固有名詞の分類性能がかなり高かった．

表 2　カテゴリ分類結果の適合率

カテゴリ	普通名詞	固有名詞	Ave.
上位語	0.43	0.68	0.56
属性語	0.59	0.79	0.69
Ave.	0.51	0.74	0.63

5．考察

評価結果に対する考察を行なう．即応性評価において，汎用辞書ではほとんどエントリが見つけられなかった．これは，注目キーワードの多くが「龍馬伝」「竹内結子」「スリーエフ」といった固有名詞であったことが原因である．逆に，Wikipediaの場合は新語や固有名詞であっても比較的短時間のうちにページがエントリされる特性が即応性確保に繋がっていると思われる．

提案手法についてみると，Wikipediaには及ばないものの，汎用辞書を大きく上回る性能が得られた．このことは，提案手法が静的な対応が困難な新語や固有名詞に対して有効であることを示している．

また，対応できなかったケースのほとんどは，適切なデスクリプタがある程度獲得できているが「語の素描」として不十分な場合であった．hit ratio(デスクリプタが獲得できた数)で比較した場合，提案手法とWikipediaと同等の性能となることから，今後，獲得するデスクリプタ数を確保できれば，提案手法がWikipediaに匹敵する性能を確保できる可能性を示唆するものである．

「擁立」「FOIL」「ルックアップ」などはWikipediaでは対応できなかったが，提案手法では対応できた，あるいはデスクリプタを獲得できたケースである．「擁立」は普通名詞であり，汎用的過ぎることがWikipediaにエントリされなかった原因であると思われる．「FOIL」や「ルックアップ」は様々な手法や形式の短縮

語であったり，多義で使われている．さらに，個々の意味で利用される語自体の認知度が低く，Wikipediaへエントリされるに至らないことが原因と考えられる．現時点では完全ではないが，提案手法であれば普通名詞，固有名詞を問わず対応可能であることがわかった．

ランキング性能については，平均MRR値は0.74であった．この結果から，平均的に上位3位以内に適切なデスクリプタを提示できることがわかる．表3は適切なデスクリプタ集合の例(上位13位)であり，人手評価による結果を適切(○，△)，不適切(×)で示している．この結果からも，上位ランキングのデスクリプタは概ね適切と評価されており，MRRによる評価結果を支持するものである．

表3 デスクリプタ集合の例(1)

クエリ	デスクリプタ	*score*	評価
龍馬伝	1年間のドラマ	0.271	○
	時代劇番組	0.123	○
	ウジウジウダウダ	0.097	×
	世界	0.093	△
	誠実	0.081	△
	ぐるぐるパーマ	0.062	△
	NHK の大河ドラマ	0.062	○
	勢い	0.043	△
	危惧	0.039	×
	自分	0.039	×
	封建的暗黒社会	0.034	△
	骨太のドラマ	0.028	△
	ドラマ	0.028	○
竹内結子	別れ方	0.303	○
	美人先生	0.175	○
	女優	0.153	○
	美人	0.149	○
	産後ダイエット成功者	0.039	○
	女性	0.033	○
	美男美女だらけ	0.011	○
	嫁	0.011	○
	超弩級の美人	0.010	○
	嗚咽	0.008	○
	イキイキ	0.008	○
	石田ゆり子	0.007	○
	嫁さん	0.006	○
	：	：	
地震速報	情報	0.225	○
	テロップ	0.219	○
	客観データ	0.097	○
	緊急	0.093	○
	プッシュ型の通信	0.077	○
	チャイム	0.075	○
	リアルタイム情報	0.043	○
	速報性	0.022	○
	命	0.021	○
	ピンポン	0.021	○
	データ	0.019	○
	災害の時の緊急速報	0.017	○
	ニュース的	0.013	○
	：	：	
iPad	端末の市場	0.264	△
	端末	0.198	○
	タブレット	0.087	○
	デバイス	0.068	○
	タッチスクリーン式端末	0.026	○
	メディアタブレット	0.023	○
	タッチ	0.023	○
	デバイスの普及	0.023	△
	電子書籍	0.023	○
	情報端末	0.022	○
	タブレット型	0.020	○
	タブレットデバイス	0.015	○
	製品	0.014	○
	：	：	

図10は，妥当と判断されたデスクリプタの順位の度数を示したものである．この結果をみると，順位1位で妥当なデスクリプタを提示できたケースが多い．さらに，ほとんどの場合順位3位までに妥当なデスクリプタを獲得していることも確認できる．

表4は，有効な結果が得られなかった例である．これらの中には，デスクリプタは獲得できたが妥当なものが得られなかったケースと，デスクリプタ自体が十分に獲得できないケースがあった．

前者については，クエリの抽象度が高過ぎるため，それ自体を表現するデスクリプタがそもそも想像しにくいことが原因であると思われる．表では「紫」「ペン回し」が該当する．例えば「紫」は，ランク1位のデスクリプタ「色」は上位概念として解釈可能であるが，その他のデスクリプタ「ピンク，緑，青，‥」は類義語である．類義語自体はクエリ語と関連の強い要素であるが，クエリ語を素描する要素としては必ずしも適切とはいえない．

後者については，クエリの実体の認知度が低すぎることが影響しており，「ラミノーズテトラ」「ベンター博士」が該当する．これらのクエリ語は，ごく一部の愛好家や専門家以外には知られていないため，クエリ自体がWeb文書中に記述され難い．よって，参照する文書範囲を拡大したとしても十分なデスクリプタを確保できる見込みはほとんど期待できない．

上記クエリはデスクリプタ集合による素描は困難である．これらのクエリについては，画像を利用するアプローチなどが効果的であると思われる．

表4 デスクリプタ集合の例(2)

クエリ	デスクリプタ	*score*	適切性
紫	色	0.43	○
	ピンク	0.12	×
	緑	0.07	×
	青	0.06	×
	茶色	0.02	×
	グレー	.023	×
	黒	.021	×
	微妙	.017	×
	ブルー	.017	×
	紺	.017	×
	：	：	
ペン回し	回転運動	0.22	○
	講座サイト	0.18	×
	思考上	0.18	×
	支点	0.14	○
	自転	0.12	×
	余計	0.074	×
	普通	0.042	×
	技術	0.040	△
ラミノーズテトラ	体型	1.00	○
ベンター博士	戦略的人材	1.00	○

さらに，全体として有効な結果が得られていても，適切とはいえないものが含まれたり，同義語，類義語が上位に複数ランキングされるケースもあった．例えば「iPad」の場合，総体としては有効といえるが，1位のデスクリプタが「端末の市場」であり最適とはいいにくい．その他，「タブレット」「メディアタブレット」「タブレット型」などの同義語や類義語が上位に複数提示されており，クエリ語の総体を把握する目的を考えると効率が悪い．

この問題に対処するためには，スコア計算モデルの精緻化，フィードバック処理の階層化，ランキングを実施する直前での同義語や類義語に対する同一性判定処理の導入，分類されたカテゴリ毎にランキングを行なうなどしてランキング対象を削減する方法などが考えられる．

スコア計算モデルに関しては，現状の計算モデルは低頻度が過剰に強調される傾向を持つことがわかっている．この傾向を補正することにより，不適切なデスクリプタに高いスコアが付与される可能性を押さえることができると思われる．

フィードバック処理については，現状ではSTEP5〜STEP7において一回のフィードバックによりデスクリプタが決定されている．これに対して，複合的フィードバック処理を設定することにより，デスクリプタの決定精度を高度化することも考えられる．現在，調査に基づき比較的高頻度で出現する指標表現 $\omega=\{$のように，に似た，みたいな$\}$を対象とした複合フィードバック処理への効果を検討中である．

同一性判定については，これによって同義デスクリプタが統一されるので，ランキング上位におけるデスクリプタの意味的重複を避けることに繋がる．分類結果毎にランキングを行なうことで，特定種類のデスクリプタが上位ランクに集中するような偏りを抑制することができる．

次に，分類性能について考察する．全体的には0.63という成功率が得られ，一定の有効性は確認できた．

普通名詞，固有名詞，上位語，属性語のそれぞれについてどのような違いがあるかをみるために，各項目間の組み合わせ（(A1)普通名詞における上位語と属性語の分類結果，(A2)固有名詞における上位語と属性語の分類結果，(B1)上位語における普通名詞と固有名詞の分類結果，(B2)属性語における普通名詞と固有名詞の分類結果）について，各項目の度数に基づいてフィッシャーの正確確率検定を実施した．その結果，(B1)にのみ5％有意水準で有意差がみられた．このことは，上位語への分類に関しては固有名詞の方が明らかに分類性能が高いことを意味しており，固有名詞の分類性能向上に寄与しているのは上位語への分類結果であることがわかる．

このことについて考察すると，まず固有名詞では，対象が具体的で一意に決まることがほとんどであるが，普通名詞には多義語や抽象的な語も多い．そのため，普通名詞のデスクリプタの分類性能を下げる主要因となる．一方，属性語への分類性能が低かった要因としては，3章(STEP9)で述べた第三ステップでの「XのY」を用いた判定処理において，定型パターンが有効に機能しておらず，誤判定が多かったことが挙げられる．これについては，今回閾値の設定が厳密でなかった定型パターンの拡充による厳密化が必要であると考えている．

その他，判別誤りを招く他の要因として，実世界の動向の影響が考えられる．データ収集時期直前に突発的な事件やイベントが発生した場合に，獲得されるデスクリプタに偏りが生じる場合がある．例えば，オリンピック開催直後の競技名や選手名，リコール問題が発生した直後の製品メーカ名などが挙げられる．このような場合，一時的に上位にランクされたデスクリプタは評価結果を下げる要因となる．

この問題については，分類結果を蓄積し，ある程度長期に渡る分類結果を総合して最終判定を行なうことを考えている．まず，分類ログを用いて，各クラスに属するデスクリプタ集合の分布を求める．次に，新たに取得したデスクリプタ分類結果と上記分類結果を比較し，ズレが生じている場合は一時的な偏りが生じていると推測し，該当するデスクリプタを他と区別して表示することで過去の分類結果を利用しながら動的処理を実現できる．ただし，これらは突発的事象やイベントなどの動向情報を反映したバーストワードであると見ることもでき，この性質を利用した語の認識変化を追跡する応用可能性も考えられる．

6．おわりに

本論文では，情報要求に対して文章や名詞句で回答する代わりに，関連する知識断片を網羅的に収集提示して「語の意味」を素描する手法を提案した．提案手法は，比喩検出のメカニズムを応用することにより，WWWから語の意味を素描する断片知識を効率よく獲得・整理して提示し，ユーザに連想的な理解を促す．

さらに，本手法の実装システムMurasakiを用いて，その抽出性能およびランキング性能，分類性能に関する評価実験を実施し，その結果から，提案手法が有効であることを確認した．今後は，デスクリプタ獲得性能の精緻化を進めるとともに，デスクリプタの同一性判定と分類性能の精緻化を予定している．特に分類性

能については，複数の定型パターンや並列構造を利用した方法[24]と，比喩性判定のメカニズム[25]を併用することを考えている．

また応用として，分類機能を利用した情報検索のキーワード拡張や，non-factoid型質問応答や検索意図を考慮した情報検索，人間判断をフィードバック情報として利用するインタラクティブな比喩的素描システムの構築などにも取り組みたい．

謝辞

本研究は，科学研究費補助金(基盤研究(C)20500833)の助成を受けています．

参 考 文 献

[1] 村田真樹，内山将夫，井佐原均："類似度に基づく推論を用いた質問応答システム"，情報処理学会研究報告，NL135-24，pp.37-42，2001.

[2] 佐々木裕，磯崎秀樹，平博順，広田啓一，賀沢秀人，平尾努，中島浩之，加藤恒昭："質問応答システムの比較と評価"，電子情報通信学会技術報告，NLC2000-24，pp.17-24，2000.

[3] Ellen M. Voorhees : "The Evaluation of Question Answering Systems : Lessons Learned from the TREC QA Track", *Proceedings of the LREC 2002 Workshop on Question Answering－Strategy and Resources*, pp.1-4, 2002.

[4] Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui : "Question Answering Challenge (QAC1) An Evaluation of QA Tasks at the NTCIR Workshop 3", *Papers from the 2003 AAAI Spring Symposium "New Directions in Question Answering"*, pp.1-4, 2003.

[5] 諸岡心，福本淳一："Why型質問応答のための回答選択手法"，電子情報通信学会技術報告（テキスト情報の要約と掲示に関わる自然言語処理シンポジウム），NLC2005-107，pp.7-12，2006.

[6] Tsuneaki Kato, Junichi Fukumoto, Fumito Masui, and Noriko Kando : "Are open-domain question answering technologies useful for information access dialog?", *ACM Transactions on Asian Language Information Processing*, Vol.4, No.3, pp.243-262, 2005.

[7] Ellen M. Voorhees and Dawn M. Tice : "The TREC-8 Question Answering Track", *Proceedings of LREC2000*, pp.1501-1508, 2000.

[8] 奥村学，難波英嗣："テキスト自動要約"，オーム社，2005.

[9] 桝井文人，福本淳一，荒木健治："比喩解釈を目的とするWorld Wide Webを利用した属性値の適合性判定とそのフィードバック"，電子情報通信学会誌，Vol. J89-D，No.9，pp. 860-870，2006.

[10] Nancy Chinchor : "Template Element Task Object Definition Version 4.2", *MUC-7 Informtation Extraction Task Definition*, pp.9-21, 1998.

[11] 芳賀純，子安増生："メタファーの心理学"，誠信書房，1990.

[12] G. Lakoff and M. Johnson : "*Metaphors We Live by*", The University of Chicago Press, Chicago, IL, 1980.

[13] 中村明："比喩表現の理論と分類"，共立出版，1977.

[14] Marti A. Hearst : "Automatic Acquisition of Hyponyms from Large Text Corpora", Proceedings of *14th International Conference on Computational Linguistic(COLING92)*, pp.539-545, 1992.

[15] Marti A. Hearst : "Automated Discovery of WordNet Relations", Christiane Fellbaum, editor, *WordNet : An Electronic Lexical Database(Chapter 5)*, The MIT Press, 1998.

[16] Matthew Berland and Eugene Charniak : "Finding parts in very large corpora", *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp.57-64, 1999.

[17] 山田一郎，住吉英樹，柴田正啓："ニュース記事に出現する用語と説明文の意味関係自動獲得"，情報処理学会研究報告，NL152-21，pp.145-152，2002.

[18] 安藤まや，関根聡，石崎俊："定型表現を利用した新聞記事からの下位概念単語の自動抽出"，情報処理学会研究報告，2003-NL-157-11（2003-FI-72），pp.77-157，2003.

[19] Patrick Pantel and M. Pennacchiptti : "Espresso : Leveraging Generic Patterns for Automatically Harvesting Semantic Relations", *Proceedings of the 21th International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp.113-120, 2006.

[20] Keiji Shinzato and Kentaro Torisawa :"Acquiring Hyponymy Relations from Web Documents", *Proceedings of the 20th International Conference on Computational Linguistics*, pp.73-80, 2004.

[21] E. M. Keen : "*The SMART retrieval system-Experiments in automatic document processing*", Prentice Hall, Inc., 1971.

[22] Patrick Pantel, Deepak Ravichandran, and Eduard Hovy : "Towards terascale knowledge acquisition", *Proceedings of the 20th international conference on Computational Linguistics (COL*ING-04), pp.771-777, 2004.

[23] 株式会社日本電子化辞書研究所："EDR電子化辞書仕様説明書"，株式会社日本電子化辞書研究所，1993.

[24] Zornitsa Kozareva, Eduard Hovy, and Ellen Riloff", Learning and Evaluating the Content and Structure of a Term Taxonomy", *Papers from the AAAI Spring Symposium*, SS-09-07, pp.50-57, 2009.

[25] 桝井文人，福本淳一，椎野努，河合敦夫："確率的尺度を用いた比喩性判定手法"，自然言語処理，Vol.9, No.5，pp.71-92，2002.

（2010年5月15日　受付）
（2010年10月11日　採録）

[問い合わせ先]
〒090-8507　北海道北見市公園町165
北見工業大学 工学部 情報システム工学科
桝井　文人
TEL：0157-26-9332
FAX：0157-26-9344
E-mail： f-masui@mail.kitami-it.ac.jp

著 者 紹 介

桝井 文人［正会員］

1990年岡山大学理学部地学科卒業．同年，沖電気工業（株）入社．2000年三重大学工学部情報工学科助手，2004～2005年北海道大学大学院情報科学研究科客員研究員．2009年より北見工業大学工学部情報システム工学科准教授．現在に至る．博士（工学）．質問応答，知識抽出，比喩理解などの研究に従事．電子情報通信学会，言語処理学会，人工知能学会，日本設備管理学会各会員．

木村 泰知［非会員］

2001年同大学院工学研究科修士課程修了．2004年北海道大学大学院工学研究科博士後期課程修了．北海道大学大学院情報科学研究科技術研究員を経て2005年小樽商科大学商学部助教授，2007年同准教授，現在に至る．博士（工学）．自然言語処理，特にウェブを利用した政治情報抽出の研究に従事．電子情報通信学会，情報処理学会，言語処理学会，人工知能学会，選挙学会各会員．

荒木 健治［非会員］

1982年北海道大学工学部電子工学科卒業．1988年同大学大学院工学研究科電子工学専攻博士後期課程修了．同年，北海学園大学工学部電子情報工学科助手．1989年同講師，1991年同助教授，1998年同教授．1992-1993年，2000年米国スタンフォード大学CSLI客員研究員．1998年北海道大学大学院工学研究科電子情報工学専攻助教授．2002年同教授を経て2004年北海道大学大学院情報科学研究科メディアネットワーク専攻教授，現在に至る．工学博士．機械翻訳，音声対話処理などの研究に従事．電子情報通信学会，情報処理学会，人工知能学会，言語処理学会，日本認知科学会，ACL，IEEE各会員．

ジェプカ・ラファウ［非会員］

1996年アダムミツキェヴィッチ大学新言語学部卒業，1999年同大学新言語学部修士．2004年北海道大学大学院工学研究科博士後期課程修了．同年北海道大学大学院情報科学研究科助手．2008年より同助教，現在に至る．工学博士．感情処理の常識的知識抽出，感情理解，機械倫理の研究に従事．情報処理学会，言語処理学会，人工知能学会各会員．

福本 淳一［非会員］

1984年広島大学工学部第2類（電気系）卒業．1986年同大学大学院工学研究科システム工学専攻博士前期課程修了．同年沖電気工業（株）入社．1992-94年英国マンチェスター科学技術大学言語学部Ph.D.コース在学．2000年立命館大学理工学部助教授．2004年～2005年米国USC/ISI客員研究員．2006年立命館大学情報理工学部メディア情報学科教授，現在に至る．Ph.D.質問応答，情報抽出，談話構造解析の研究などに従事．電子情報通信学会，情報処理学会，言語処理学会，人工知能学会，ACL各会員．

WWW‐based Figurative Descriptions for Japanese Word

by

Fumito MASUI, Rafal RZEPKA, Yasutomo KIMURA, Jun‐ichi FUKUMOTO and Kenji ARAKI

Abstract：

In this paper, we propose a method for describing a Japanese word, not with explaining or defining sentences, but with figurative descriptions. Utilizing a simile pattern, our method gathers a large number of noun‐noun relations from the World Wide Web. On the basis of those relations and their statistical information, associative pieces of knowledge called descriptors are estimated. The descriptors, which describe a query word figuratively, are sorted by ranking in order of descriptive ability level with generality and locality. Moreover, combining property of figurative relation and some fixed patterns, the descriptors are classified into concept words, attribute words, and the others. As output, a set of sorted descriptors is shown with several types of output forms.

Some experiments using a prototype system "Murasaki" have been conducted. The experimental results show that the fundamental performance of our method is significantly better than the bag‐of‐words approach. Additionally, the responsiveness for hot keywords on information retrieval web sites shows that the outcome of the evaluation had 60% precision, which exceeds that of a common dictionary. The method also functioned effectively in ranking performance (74% on MRR) and classification performance (63% accuracy). Furthermore, it is possible that the proposed method could be comparable to Wikipedia if steady coverage of the figurative descriptions for a query word could be ensured.

Keywords：figurative descriptions, descriptor, simile expression, fixed pattern

Contact Address：**Fumito MASUI**

Department of Computer Science, Kitami Institute of Technology
165, Kouen‐cho, Kitami, 090‐8507, JAPAN
TEL ：0157‐26‐9332
FAX ：0157‐26‐9344
E‐mail ：f-masui@mail.kitami-it.ac.jp