Vol.2011-DBS-153 No.25
2011/11/3

説明からの用語検索のための緩和によるクエリ生成とページ中の位置を考慮した用語抽出

木場 由布子[†1] 湯本 高行[†1]
新居 学[†1] 高橋 豊[†1]

我々はユーザの説明から Web を用いて用語を検索する手法を考案している．本手法では，Web を用いてユーザの説明からクエリを生成し，Web を用いて逆引きを行い用語を抽出した後，順引きを行い用語を検証する．このとき，用語がユーザの説明に適しているかどうかを検証することは比較的容易であるが，ユーザの知りたい用語を漏れがなく抽出し，かつ用語の候補を絞り込むことは難しい．そこで本稿では，ユーザの説明を緩和することでクエリを生成し，得られた Web ページ中の用語の位置を考慮して尤もらしい用語を優先的に選択する手法を提案する．実験より，クエリを緩和することで正解となる用語の出現率を向上させることができた．また，Web ページ中のタイトルから用語を優先的に抽出することが有効であることがわかった．

Relaxed Query Generation and Position-based Term Extraction for Term Search from User's Description

Yuko Koba,[†1] Takayuki Yumoto,[†1] Manabu Nii[†1]
and Yutaka Takahashi[†1]

We are developing a method for finding an object name corresponding to a description given by a user from the Web. Our method consists of three phases, the query generation phase, the term extraction phase and the term validation phase. We have already achieved enough performance in the term validation phase. In the query generation phase and the term extraction phase, however, it is still difficult to find all of possible candidates with reducing the number of them. So, we propose two algorithm of relaxed query generation and position-based term extraction. The precision by our query generation algorithm was higher than the precision when we used queries consisting of all the words in the description. And we found that it is efficient to extract term from the title of web pages.

1. はじめに

日常生活において，説明はできるが名前がわからない場合がある．例えば，旅行先で見た鳥の名前であったり，テレビコマーシャルに出演している人の名前を知りたい場合である．このようなとき，辞書や Web から知りたい “用語” を探すことは難しい．何故なら，ユーザの説明がその用語にとって主要な説明であるとは限らないからである．つまり，その説明が曖昧であったり，さらには不正確であったりする可能性がある．そこで我々は，ユーザが知りたい用語に関する説明を入力することで，知りたい用語を Web を用いて検索する手法を提案している．Web にはさまざまな情報が記述されているため，Web を用いることで曖昧な説明や不正確な説明に対応できると考えている．本手法では，Web を用いてユーザの説明から用語を逆引きした後，Web を用いて用語の説明を順引きし，用語がユーザの説明に適しているかどうかを検証する．具体的には，ユーザの説明のキーワードを組み合わせることでクエリを生成する．次に，クエリを用いた Web の検索結果から用語を抽出する．そして，抽出した用語自身をクエリとして用いた検索結果とユーザの説明を比較することで用語を検証する．このとき，ユーザの説明と用語の検索結果を比較することによって用語がユーザの説明に適しているかどうかを検証することは比較的容易であるが，ユーザの知りたい用語を漏れがなく抽出し，かつ用語の候補を絞り込むことは難しい．何故なら，クエリを用いた検索結果に知りたい用語がそもそも存在していなかったり，存在していたとしても，単純にすべての用語を抽出すれば，数が多すぎるからである．また候補数が多すぎる場合，用語を検証する際にすべての用語を一度に検索することは現実的でない．そこで本稿では，ユーザの説明を緩和することでクエリを生成し，クエリを用いて得られた Web ページ中の用語の位置を考慮して尤もらしい用語を優先的に選択する手法を提案する．

2. 本研究の位置づけ

2.1 本研究の概要

我々はユーザの説明から Web を用いて用語を検索することを目的としている．本研究では，説明は用語の種類とそれを形容する部分から成り立つと考える．そこで，この用語の種類をクラス，形容部分を説明句と定義する．ユーザの説明の例を図 1 に示す．図 1 は，“ア

†1 兵庫県立大学大学院工学研究科
Graduate School of Engineering, University of Hyogo

Vol.2011-DBS-153 No.25
2011/11/3

餌を週一回くらい少しやるだけでもう三年も生きている魚

図 1 アカヒレの説明文

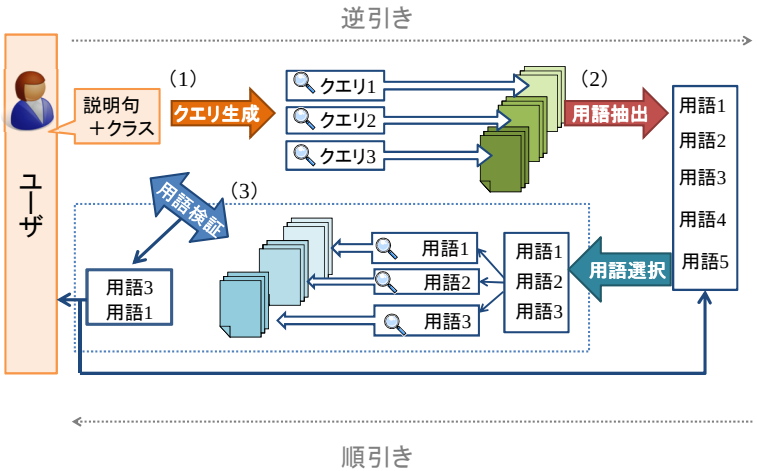

図 2 提案手法の全体像

カヒレ”という用語を説明した文であり，クラスは “魚”，説明句は “餌を週一回くらい少しやるだけでもう三年も生きている”となる．

以下，ユーザの説明句とクラスを入力として，Web を用いて用語を検索する手法について述べる．本手法では，ユーザの説明を入力として，Web に対して逆引きを行うことで用語を抽出し，抽出した用語自身を Web に対して順引きを行うことで用語を検証する．図 2 に手法の全体像を示す． 図 2 に示すとおり，提案手法は以下の 3 段階で成り立っている．

(1) ユーザの説明からのクエリ生成
(2) Web からの用語抽出
(3) ユーザの説明と用語の検索結果の比較による用語検証

クエリ生成ではクラスと説明句のキーワード集合の組合せによりクエリ集合を生成する．用語抽出ではクエリ集合を用いた Web の検索結果集合から用語を抽出し，用語の Web ページ中の位置を考慮して，用語自身にスコアを設けておく．抽出したすべての用語から，用語のスコアを用いていくつかの用語を選択し，その用語が妥当であるかどうかを検証する．用語検証では，抽出した用語をクエリとした Web の検索結果とユーザの説明を比較する．検証を終えた用語から知りたい用語が見つからなかった場合は，新たに異なる用語をいくつか選択し，それらの用語を検証するという過程を繰り返すことで用語を検索していく．なお，(3) の用語検証に関しては，すでに提案した手法[1] を用いるため割愛し，本稿ではクエリ生成及び用語抽出について述べる．3 章にクエリ生成手法を示し，4 章に用語抽出手法を示す．

2.2 関 連 研 究

説明から用語を検索することを実現するサービスとして，知識サーチエンジンである Powerset[2] や WolframAlpha[3] がある．しかし，これらのサービスにおける入力となる説明は大概の人が知っている明瞭な説明でなければ知りたい用語を検索することができず，本研究で対象としている，不正確または曖昧な説明から用語を検索することは困難である．一方で，不正確または曖昧な説明からでも用語を検索することができるコミュニティQA サイト（“ 以下 CQA サイト ”とする）というサービスが存在する．CQA サイトは，質問者がある特定のことを知りたかったり，他の人にアドバイスを得たい場合に使用する．このサービスは，あるユーザが投稿した質問に他のユーザが回答するシステムである．そして，質問者がその回答を適切であると判断した場合，良回答とする．質問が曖昧または不正確に記述されていても，人が質問を理解することで質問者は知りたい回答を得ることができる場合がある．しかし，本研究では短時間で用語を検索することを目指すのに対して，CQA サイトでは人が回答するゆえ即座に知りたい用語が得られるとは限らない．この問題に対して Watson プロジェクト[4] では答えが一意に決まる説明である質問に対して短時間かつ高確率で解を得るという結果を残している．しかし，本手法では，Web を対象として特定のデータベースを生成せずに用語を検索するのに対し，Watson プロジェクト[4] ではあらかじめ索引を用いた特定のデータベースを生成しなければならない．本研究と同様，うろ覚えで思い出せない用語を検索することができる稲川ら[5] の研究でも，用語を特定するキーワードの索引を事前に Web ページから抽出し，特定のデータベースを生成しており，特定のデータベースを生成せずに Web を対象として用語を検索する本手法とは異なる．

また，キーワードを緩和する検索手法として金子ら[6] はユーザのクエリをユーザの意図による緩和の度合いに沿って変換してクエリを生成することで，ユーザの意図に沿った Web のランキング結果を得ている．しかし，本研究ではユーザの意図に沿った Web のランキング結果を得ることが目的ではなく，用語を検索することを目的としてそれに合ったクエリ生成手法を提案している．また，Kumaran[7] らはクエリの質を解析することによって，長い

Vol.2011-DBS-153 No.25
2011/11/3

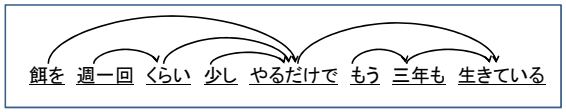

図 3 “餌を週一回少しやるだけでもう三年も生きている”という文の係り受け解析結果

記述から無用な語を省き，良質なクエリを生成している．しかし，Kumaran[7] らの研究ではあらかじめクエリの学習データが必要となり，本手法ではユーザの説明のみからクエリを生成するため，その点で異なる．

本手法ではクエリとの距離を用いて用語のランキングを行うが，クエリとの距離を用いる手法として，望月ら[8] は文書中の文脈情報である語彙的連鎖（意味的なまとまり）を利用し，検索要求であるクエリと強く関連している部分を検索するパッセージ検索手法を提案している．

3. 説明の緩和によるクエリ生成

ユーザの説明からクエリを生成する手法について述べる．ユーザの説明からクエリを生成する場合，最も単純な方法はユーザの説明のすべてのキーワードをクエリに使用することであるが，キーワードが多すぎると一般的にクエリに該当する Web ページ自体が存在しないことがあったり，正解となる用語が存在する Web ページを得ることが難しい．また，キーワードを減らしてクエリを生成する場合，最も単純な方法はクエリのキーワードのすべての組合せを生成することであるが，組合せ爆発が起こり，実用的ではない．そこで，本手法では正解となる用語の Web ページにおける出現率を効率よく向上させるために，係り受けの関係を用いて説明句を緩和することでクエリを生成する．図 3 に “餌を週一回少しやるだけでもう三年も生きている”という文を係り受け解析した例を示す．図 3 の矢印はそれぞれ係り受け関係が成り立っている部分を指し，矢印元は係り元，矢印の先は係り先となっている． 図 3 のように，説明句を係り受け解析したとき，係り先を辿ることによって得られる文節の極大の組合せから成るフレーズのキーワード集合 $RelaxPhraseKeys_i$ およびクラスのキーワード c を組み合わせることによってクエリ q を生成する．係り先でつながっている部分を得ることで，意味のある固まりを得ることができると考える．

以下に “餌を週一回少しやるだけでもう三年も生きている”という文を緩和して得ることができるフレーズを示す．

- 餌をやるだけで生きている
- 週一回くらいやるだけで生きている
- 少しやるだけで生きている
- もう三年も生きている

それぞれ生成するクエリ集合 Q を (1) 式に示す．

$$Q = \{q | q = \{c\} \cup RelaxPhraseKeys_i, 1 \leq i \leq |RelaxPhraseKeys|\} \tag{1}$$

$|RelaxPhraseKeys|$ は説明句を緩和して得ることができるフレーズの数を示す．なお，係り受け解析には CaboCha[⋆1] を用いる．これらのクエリ集合 Q から (2) 式のように Web ページ集合 $P(Q)$ を取得する．

$$P(Q) = \bigcup_{q \in Q} P(q) \tag{2}$$

$P(q)$ はクエリ q を用いて得られる Web ページ集合である．なおクエリには，名詞，形容詞，形容動詞，動詞のキーワードを使用する．

4. ページ中の位置を考慮した用語抽出

ユーザの説明から生成したクエリ集合 Q を用いたそれぞれの Web の検索結果の上位 N 件から用語を抽出する手法について述べる．以下に手順を示す．

(1) クエリ集合の各クエリに対して Web ページから用語 t を抽出する．
(2) 各 Web ページに対して用語 t のスコア付けを行う．
(3) クエリ集合から抽出されたすべての用語に対してスコア付けを行い，スコアが上位 X 件の用語を選択する．

ここで，用語抽出では形態素解析を行っており，本手法では MeCab[⋆2] を使用している．用語は固有名詞及び複合名詞を抽出することを目指す．辞書に発音が正しく登録されていない名詞は固有名詞である可能性が高いため，MeCab において発音が “*” と表記される単一名詞を抽出する．複合名詞に関しては，品詞が名詞で 2 語以上連結している名詞列を抽出する．

すべてのクエリ集合 Q に対する用語 t のスコア $score(t, Q)$ は，(3) 式のように各クエリ q で抽出された用語 t のそれぞれのスコアで最も数値が高いスコアとする．

$$score(t, Q) = \max_{q \in Q} score(t, q) \tag{3}$$

⋆1 http://chasen.org/~taku/software/cabocha/
⋆2 http://mecab.sourceforge.net/

このとき，クエリ q に対する用語 t のスコア $score(t,q)$ は，(4) 式のようにクエリ q を用いて得られた Web ページ集合 $P(q)$ において各ページ p で抽出された用語 t のそれぞれのスコアで最も数値が高いスコアとする．

$$score(t,q) = \max_{p \in P(q)} score(t,p) \tag{4}$$

なお，Web ページ p で抽出された用語 t のスコア $score(t,p)$ の算出方法に関しては 4.1 節，4.2 節，4.3 節にてそれぞれ後述する．

最後に，スコアが上位 X 件の用語を選択し，検証する．知りたい用語が見つからなければ，さらにスコアの順位が上位 X 件の用語を選択し，すでに検証した用語以外の用語を検証していくことを繰り返す．

本稿では，以下のような用語を優先的に選択することを考えて，用語抽出においてそれぞれの手法について実験する．

手法 (A) Web ページのタイトルに存在する用語を優先的に抽出

手法 (B) Web ページの本文全体において重要な用語を優先的に抽出

手法 (C) Web ページの本文の一部において重要な用語を優先的に抽出

クエリの内容が “餌をやるだけで生きている魚”であり，知りたい用語が “アカヒレ”の場合を考える．手法（A）においては，例えば Web ページのタイトルが “餌をやるだけで生きている魚，アカヒレ”である場合など，クエリの内容がタイトルに存在し，かつ用語もタイトルに存在する場合などが考えられる．さらに，Web ページのタイトルが “アカヒレ”である場合は，クエリの内容が本文の一部に一致し，かつ用語がタイトルに存在する場合などが考えられる．手法（B）においては，Web ページのタイトルが “餌をやるだけで生きている魚って何”である場合にクエリの内容がタイトルに存在し，本文全体において “メダカ”などと同様に “アカヒレ”という用語が頻繁に出現した場合などが考えられる．手法（C）においては，Web ページのタイトルが “魚図鑑”である場合にクエリの内容が本文の一部に一致し，その近傍に “アカヒレ”が存在する場合などが考えられる．以下，それぞれの用語抽出手法及び，その手法に相当する各 Web ページに対する用語 t のスコア $score(t,p)$ の算出方法を示していく．

4.1 タイトルからの用語抽出

Web ページ p のタイトル $title(p)$ から用語を抽出する．この場合，クエリ q との一致率が大きいページのタイトルほど重要であると考え，クエリ q を用いて得られた Web ページのキーワード集合 $keyword(p,q)$ 中にクエリ q 内のキーワードがどれだけ含まれているかの割合で用語 t にスコア付けを行う．

$$score_{title}(t,q,p) = \frac{|q \cap keyword(p,q)|}{|q|} \tag{5}$$

このキーワード集合にはクエリと同じく名詞，動詞，形容詞，形容動詞を用いる．(4) 式における $score(t,p)$ は，タイトルから用語を抽出する場合は $score_{title}(t,q,p)$ を用いる．

4.2 本文全体からの用語抽出

Web ページの本文 $content(p)$ の全体から用語を抽出する．本文全体において重要な順番にランキングを行うために，出現頻度が高い用語ほどスコアは高くなるように，同じ Web ページ p から抽出された用語 t を出現頻度の降順で並べた時ときの順位の逆数である RR を用いてスコア付けを行う．

$$score_{tf}(t,p) = RR(t,p) \tag{6}$$

(4) 式における $score(t,p)$ は，本文全体から用語を抽出する場合は $score_{tf}(t,p)$ を用いる．

4.3 本文の一部からの用語抽出

Web ページの本文の一部から用語を抽出する手法について述べる．まず，Web ページの本文 $content(p)$ の全体から用語を抽出する．次に，本文においてクエリの近傍が重要であると考えるため，用語とクエリのキーワードの出現位置を用いて (7) 式のように用語にスコア付けを行う．

$$score_{neighbor}(t,q,p) = \begin{cases} \dfrac{1}{|K|} \displaystyle\sum_{k \in K} \dfrac{1}{\log(dist(t,k,p) \ +2)} & (|K|>0) \\ 0 & (otherwise) \end{cases} \tag{7}$$

(7) 式では，Web ページ p 内における用語 t とクエリ q に含まれているキーワード k との距離 $dist(t,k,p)$ が近い程スコアが高くなるようにスコア付けを行う．また，キーワード集合 K は，クエリに含まれるキーワードでかつ，タイトルに含まれる用語は本文全体に均一に関係していると考えるため，タイトルに含まれていないキーワード集合とする．$|K|$ はキーワード集合の数であり，Web ページ p にキーワード k が存在しなければ用語のスコア $score_{neighbor}(t,q,p)$ の値を 0 とする．キーワード集合 K を（8）式に示す．

$$K = q \setminus title(p) \tag{8}$$

用語 t とキーワード k との距離 $dist(t,k,p)$ は，それぞれが存在する文の単位の距離であり，Web ページ p に複数存在する場合はキーワード間での最も近い距離を使用する．図 4 に距離 $dist(t,k,p)$ の考え方の例を示す．図 4 は，1 行が 1 文を示しており，上からそれぞ

Vol.2011-DBS-153 No.25
2011/11/3

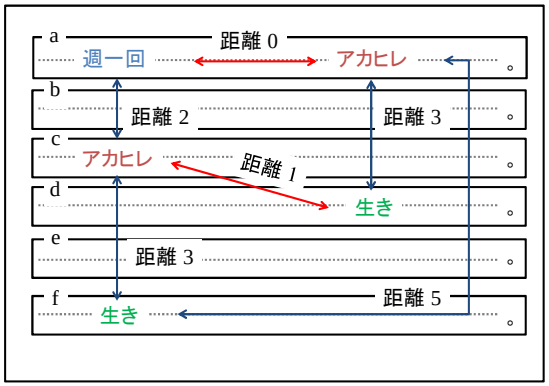

図 4 距離の考え方の例

れ文 a，文 b，文 c，文 d，文 e，文 f を示している．用語が “アカヒレ”，クエリに含まれるキーワードが “週一回”である場合を考える．“アカヒレ”と “週一回”では，それぞれ文 a に同時に存在している場合が文の距離が最も小さいため，これらのキーワード間の距離は 0 とする．文 c の “アカヒレ”では “週一回”との距離は 2 となり，0 より大きい値となる．次に，用語が “アカヒレ”，クエリに含まれるキーワードが “生き”である場合を考える．“アカヒレ”と “生き”では，“アカヒレ”が文 c，“生き”が文 d に存在している場合が文の距離が最も小さいため，これらのキーワード間の距離は 1 とする．“アカヒレ”が文 a，“生き”が文 d の場合は距離は 3，“アカヒレ”が文 c，“生き”が文 f の場合は距離は 3，“アカヒレ”が文 a，“生き”が文 f の場合は距離は 5 となり，1 より大きい値となる．なお，Web ページ p にクエリに含まれるキーワード k が存在しなければ用語 t との距離 $dist(t,k,p)$ は無限大になるとみなし，この場合（7）式の $\frac{1}{\log(dist(t,k,p)\ +2)}$ の部分は値を 0 とする．

（4）式における $score(t,p)$ は，本文全体から用語を抽出する場合は $score_{neighbor}(t,q,p)$ を用いる．

5. 実　　験

ユーザの説明を緩和することで，生成したクエリを用いて得られる Web ページ集合中において，正解となる用語の出現率が向上するかどうかを検証するためのクエリ生成に対する実験を行った．また，ユーザの説明から得られる用語の上位 X 件に正解用語が得られるかどうかの用語抽出に対する実験を行った．用語抽出の実験に関しては 4 章で示したタイトルから用語を抽出する手法，本文全体で重要な用語を抽出する手法，本文一部で重要な用語を抽出する手法のそれぞれに対して実験を行った．

表 1 実験に使用した質問文及び説明及びクエリの数

	件数
質問文	41
説明	107
クエリ	314

各実験では，ユーザの説明として OKWave や Yahoo 知恵袋の CQA サイト内の用語を問う質問文 41 件を使用した．正解用語は質問文に対して良回答内に示されている用語を基に著者が調査して妥当であると判断した用語を用いた．また，これら 41 件の質問文を読点や改行で区切った単位をひとつの説明とみなし，説明句とクラスから成る形式に簡易化した 107 件の説明を使用した．実験に用いる Web の検索結果取得数 N は 20 とした．なお，Web ページにおいては OKWave，Yahoo 知恵袋などの CQA サイトを除いて実験を行った．表 1 に実験に使用した質問文及び説明及びクエリの数を示す．

5.1 緩和によるクエリ生成の実験

107 件の説明を入力として，説明を緩和して生成したクエリ集合を用いた Web ページ集合内に正解用語が出現するか否かを，説明を緩和しない場合のクエリを用いた手法をベースライン手法として比較した．図 2 において，説明を入力として (1) のクエリ生成から (2) の用語抽出前の段階における Web ページ集合内での正解用語の出現率 $Recall_{answer}$ を調査した．

$$Recall_{answer} = \frac{|Exp_{answer}|}{|Exp_{user}|} \tag{9}$$

$|Exp_{answer}|$ は Web ページ集合内に正解が出現したユーザの説明の数を示し，$|Exp_{user}|$ はすべてのユーザの説明の数である 107 となる．

表 2 に各手法における Web ページ集合内での正解用語の出現率を示す．表 2 より，説明を緩和した場合の正解用語出現率は 0.60，ベースライン手法による正解用語出現率は 0.40 となり，説明を緩和してクエリを生成することで正解用語の出現率を向上させることができた．よって，ユーザの説明を緩和してクエリを生成することは有効であることがわかった．

5.2 ページ中の位置を考慮した用語抽出の実験

107 件のユーザの説明を入力として得られる用語の上位 X 件に正解用語が得られるかど

表 2 正解用語の出現率

手法	正解用語の出現率
提案手法による正解用語出現率	0.60 (64/107)
ベースライン手法による正解用語出現率	0.40 (49/107)

表 3 手法（A），手法（B），手法（C）における正解用語の MRR

	MRR
手法（A）	0.36
手法（B）	0.18
手法（C）	0.05

うか実験を行った．この実験ではユーザの説明を緩和して生成したクエリを用いた．そこで，用語を抽出した際に尤もらしい用語から順番に検証することを評価するためにパラメータ $X = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 20000, 30000, 40000, 50000$ として実験を行った．用語抽出部分に関しては，4 章に示すタイトルから用語を抽出する手法（A），本文全体で重要な用語を抽出する手法（B），本文一部で重要な用語を抽出する手法（C）に関してそれぞれ別に実験を行った．正解用語の順位の逆数の平均である MRR，正解用語の抽出率，及び各手法で抽出される用語の類似性を検証した．以下に実験結果を示す．

表 3 に手法（A），手法（B），手法（C）における正解用語のスコアの順位の逆数の平均をとった値 MRR を示す．表 3 より，MRR はタイトルから用語を抽出する手法（A）が最も高く，本文全体から用語を抽出する手法（B）が 2 番目に高く，本文一部から用語を抽出する手法（C）が最も低い値となっている．

次に，図 5 にすべての 107 件のユーザの説明に対して，それぞれ手法（A），手法（B），手法（C）においてパラメータ X にて（10）式に示すように正解用語を抽出できた精度 $Recall_{All}(X)$ を示す．

$$Recall_{All}(X) = \frac{|Exp_{method}(X)|}{|Exp_{user}|} \tag{10}$$

$|Exp_{method}(X)|$ はパラメータ X において各手法で正解が抽出できたユーザの説明の数を示し，$|Exp_{user}|$ はすべてのユーザの説明の数である 107 となる．

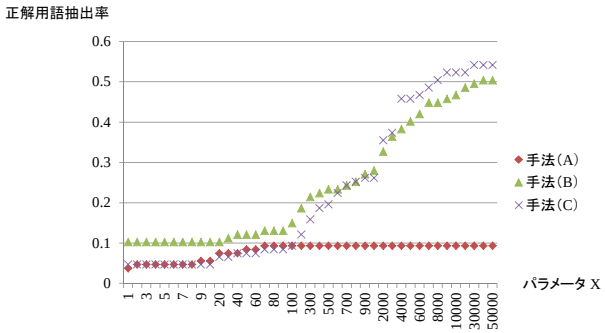

図 5 すべての説明に対する正解用語抽出精度

図 5 より，パラメータ X が 1 から 1000 の間では，本文全体から用語を抽出する手法（B）の精度が最も高いことがわかる．また，パラメータ X が 2000 から 50000 の間では本文一部から用語を抽出する手法（C）の精度が最も高いことがわかる．さらに，タイトルから用語を抽出する手法（A）は全体を通して最も精度が低いことがわかる．

次に，図 6 に各手法に対してすべての用語を抽出した際に正解用語が抽出できるユーザの説明に対して，それぞれ手法（A），手法（B），手法（C）においてパラメータ X にて（11）式に示すように正解用語を抽出できた精度 $Precision_{method}(X)$ を示す．

$$Precision_{method}(X) = \frac{|Exp_{method}(X)|}{|Exp_{method}|} \tag{11}$$

$|Exp_{method}(X)|$ はパラメータ X において各手法で正解が抽出できたユーザの説明の数を示し，$|Exp_{method}|$ はそれぞれの手法における，すべての用語を抽出した際に正解用語が抽出できるユーザの説明の数を示している．

図 6 より，タイトルから用語を抽出する手法（A）に対しては，X が 70 のときに正解用語の抽出率が 1.0 となった．このとき，タイトルにおいて抽出できる正解用語はすべて抽出できており，他の手法（B）及び手法（C）よりも正解用語の抽出率が高い．また，前半は本文全体から用語を抽出する手法（B）の方が本文一部から用語を抽出する手法（C）よりも正解用語の抽出率が高く，パラメータ X が 2000 以降は手法（B）及び手法（C）に関しては明確な差は見られない．

図 7 に各手法（A），手法（B），手法（C）における各パラメータ X で抽出される各ユー

Vol.2011-DBS-153 No.25
2011/11/3

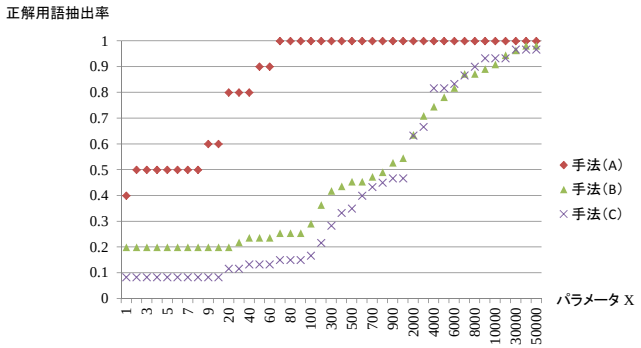

図 6 正解用語が抽出できる説明に対する正解用語抽出精度

表 4 それぞれ手法（A），手法（B），手法（C）間で用いた類似度

	類似度に用いた値
手法（A）及び手法（B）間	Simpson 係数
手法（A）及び手法（C）間	Simpson 係数
手法（B）及び手法（C）間	コサイン類似度

ザの説明に対する用語の平均類似度を示す．

タイトルから用語を抽出する手法（A）によって抽出される用語は，本文から用語を抽出する手法（B）及び手法（C）によって抽出される用語数より少ないため，手法（A）との類似性を検証する場合は Simpson 係数を使用した．本文から用語を抽出する手法（B），手法（C）間の用語の類似度に関しては，コサイン類似度を用いて算出した．(12) 式に手法（A）及び手法（B）の Simpson 係数を示す．手法（A），手法（C）間の用語の類似度に関しても同様に算出した．

$$Simpson(T_A, T_B) = \frac{|T_A \cap T_B|}{min(|T_A|, |T_B|)} \tag{12}$$

T_A，T_B はそれぞれ手法（A），手法（B）で抽出される用語集合であり，$|T_A|$，$|T_B|$ はそれぞれ手法（A），手法（B）で抽出される用語の数である．

(13) 式に手法（B）及び手法（C）のコサイン類似度を示す．T_B，T_C はそれぞれ手法（B），手法（C）で抽出される用語集合であり，$|T_B|$，$|T_C|$ はそれぞれ手法（A），手法（B）

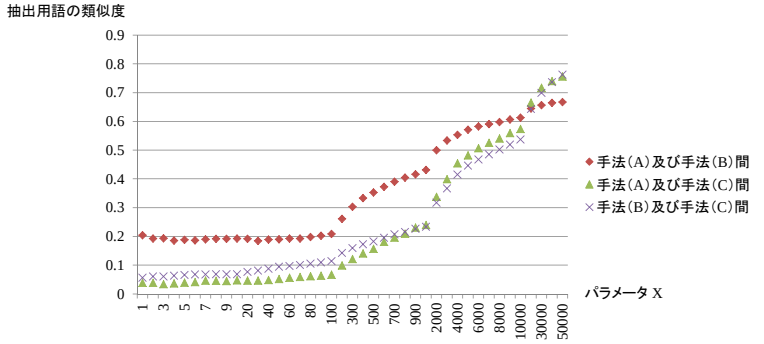

図 7 各手法において抽出される用語の平均類似度

で抽出される用語の数である．

$$Cos(T_B, T_C) = \frac{|T_B \cap T_C|}{\sqrt{|T_B||T_C|}} \tag{13}$$

図 7 より，それぞれ X が上位 100 位のときの平均類似度は手法（A）及び手法（B）間では 0.209，手法（A）及び手法（C）間では，0.0678，手法（B）及び手法（C）間においては 0.115 となり，最も高い値でも 0.209 となり，低い値となった．また，X が上位 1000 位においてもそれぞれの平均類似度は手法（A）及び手法（B）間では 0.432，手法（A）及び手法（C）間では，0.241，手法（B）及び手法（C）間においては 0.235 となり，平均類似度は 0.5 以下となり，それぞれ上位に抽出される用語は異なる傾向にあるということがわかった．

次に，各手法の上位 1 位以内で，それぞれの手法で正解用語を抽出することができた用語とそのユーザの説明を表 5，表 6，表 7 に示す．なお，本文全体から抽出する手法（B），本文一部から抽出する手法（C）に関しては，その手法でのみ上位 1 位以内で抽出できたユーザの説明とその正解用語を抜粋したものを示す．本文全体から抽出する手法（B）と本文一部から抽出する手法（C）では上位 1 位以内で正解用語を抽出することができた説明と正解用語が同じ場合も存在したが，表 6，表 7 より，異なる説明で正解用語を上位 1 位に抽出できる場合もあることが確認できた．また，タイトルから用語を抽出する手法（A）では，実際に上位 1 位では他の手法のどちらかでも上位 1 位で同じように正解用語を抽出できたが，上位 3 位以内ではタイトルから用語を抽出する手法でのみ “よく海外で観光客向けにやっている、２週間くらいで落ちるタトゥー” という説明から正解用語である “ヘナタトゥー” を抽

表 5 手法（A）で上位 1 位以内で抽出できた説明と正解用語

正解用語	説明
北岳	日本で 2 番目に高い山
TAP	SPI でも SCOA でもない筆記試験
ミサンガ	たしか、カタカナで 4 文字、1 文字目がマ行、3 文字目が「ン」か「ッ」だったような気がする輪っか
代理ミュンヒハウゼン症候群	例えばですが、消防士が、自ら火をつけ第一発見者になり消化活動にあたる症候群

表 6 手法（B）でのみ上位 1 位以内で抽出できた説明と正解用語

正解用語	説明
軍艦島	今は、コンクリートの建物ばっかりで昼間でも怖いし、不気味な島
ミサンガ	足に巻く輪っか
アオサギ	色は、白を基調とし黒やグレーの模様が入っている鳥
アカヒレ	知合いの家で見た、マグカップくらいの透明のコルク栓のビンの中で飼われていたメダカのような魚
代理ミュンヒハウゼン症候群	母親が子どもを病院へ連れて行き「お母さん　よく気がつきましたね！」などと言って貰いたくて自分の子どもを傷つける症候群

表 7 手法（C）でのみ上位 1 位以内で抽出できた説明と正解用語

正解用語	説明
エリザベスカラー	犬が手術をした後などに、傷を舐めたりしないようにするための（エリマキトカゲみたいな？）首輪
ラブアタック	上岡龍太郎さんが司会の番組

出することができていることが確認できた．よって，正解用語に関してもそれぞれ上位に抽出される用語は異なる場合があることがわかった．

以上から，用語抽出においては，全体的な正解用語の抽出精度は低いが，上位に正解用語が抽出されるタイトルから用語を抽出する手法（A）を用いて優先的に用語を抽出し，その後本文から抽出する手法（B），及び手法（C）を用いて用語を抽出するのが妥当であると考える．次に，本文全体から抽出する手法（B），及び本文一部から抽出する手法（C）に関しては，手法（B）の方が上位に正解用語が抽出される傾向があり，前半における正解用語の抽出精度は高いが，差は明確ではなく，後半においては本文一部から抽出する手法（C）の方が正解用語の抽出率は高くなることがわかった．また，手法（B），及び手法（C）において上位に抽出される用語は異なるという結果が得られたため，それぞれの手法を用いて用語を抽出することが妥当であるといえる．今後はこれらの手法を組み合わせて用語を選択する手法を具体化することが課題となる．

6. ま と め

本稿では，説明からの用語検索における手法を提案し，クエリ生成手法及び用語抽出手法を提案した．クエリ生成においてはユーザの説明を緩和しない場合の正解用語の出現率は 0.40，ユーザの説明を緩和する場合の正解用語の出現率は 0.60 となり，ユーザの説明を緩和してクエリを生成することで正解用語の出現率を向上させることができた．また，タイトルからの用語抽出手法，本文全体を重視する用語抽出手法，及び本文の一部を重視する用語抽出手法の 3 つの手法について説明からクエリを生成し，用語を抽出するまでの実験を行った．正解用語の MRR はタイトルから抽出する手法が本文から抽出する手法に比べて最も高かった．そのため，用語を選択する際に，正解用語をより正確に抽出するために，最初はタイトルから抽出された用語を検証していき，次に本文から抽出された用語を検証していくことが妥当であることがわかった．今後は，3 つの手法を組み合わせることで抽出した用語を選択する手法を具体化することが課題となる．

参 考 文 献

1) 木場 由布子，湯本 高行，新居 学，高橋 豊，説明記述からのクエリ生成による逆引き用語検索，DEIM Forum2010，D6-5 (2010).
2) Powerset, http://www.powerset.com/
3) Wolfram—Alpha-Computational Knowledge Engine,http://www.wolframalpha.com/
4) http://www-06.ibm.com/ibm/jp/lead/ideasfromibm/watson/
5) 稲川雅之，大島裕明，小山聡，田中克己，Web からの語集合間の特定関係抽出とその可視化，日本データベース学会論文誌，Vol.7，No.1，pp.175-180 (2008).
6) 金子　恭史，中村　聡史，大島　裕明，田中　克己，緩和度付き検索語の意味関連分析による検索意図推定とそのクエリ入力インタフェース，DEWS2008，B7-2 (2008).
7) G. Kumaran, and V. R. Carvalho, Reducing Long Queries Using Query Quality Predictors, Proc. of SIGIR ’09, ACM, Boston, USA, pp.564-571 (2009).
8) 望月源, 岩山真, 奥村学，語彙的連鎖に基づくパッセージ検索 ，情報処理学会，自然言語処理, Vol.6, No.3，pp.101-126 (1999).