文字列型パッセージ検索のための単語印象を考慮した語彙的言い換え

熊本　忠彦†　　田中　克己††

† 独立行政法人情報通信研究機構・自然言語グループ　〒 619–0289 京都府「けいはんな学研都市」光台 3–5
†† 京都大学大学院・情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †kuma@nict.go.jp, ††ktanaka@i.kyoto-u.ac.jp

あらまし　近年，Web 上には多種多様な情報が大量に存在しており，各種 Web 応用システムの貴重な情報源となっているが，記述スタイル（特に文体や語彙）には個人差があるため，必要な情報を正確かつ網羅的に収集するのは容易でない．現在，著者らは,「育児に参加しない父親」のようなトピック表現（文字列）がクエリとして入力されたときに，そのトピックに関して何らかの記述のある Web 文書を正確かつ網羅的に収集する検索方式の実現を目指しており，本稿では，その第 1 段階として,「育児」を「子育て」に言い換えるような語彙的言い換え方式を提案する．提案方式は, (1) クエリ中の内容語（名詞，形容詞，動詞，カタカナ）に対する言い換え候補を Web 検索により獲得する, (2) 言い換えの妥当性を 2 種類の辞書を用いて判定する, (3) 言い換え可と判定された候補を用いて新たなクエリ群を生成する，という手順からなり，言い換えの妥当性を単語どうしの前接関係・後接関係・述語関係を示す共起辞書だけでなく，ある特定の印象評価軸に沿って対比させられた 2 つの印象語群との共起関係を示す印象辞書を用いて判定する点に特徴がある．なお，本稿では 7 個のサンプルクエリを用いて提案方式を評価し，その有効性を検証する．
キーワード　クエリ展開，意見マイニング，質問応答，評判分析，文書検索，印象マイニング

Lexical Paraphrasing Using Impressions of Words for Character String-Based Passage Retrieval on the Web

Tadahiko KUMAMOTO† and Katsumi TANAKA††

† Computational Linguistics Group, National Institute of Information and Communications Technology
Hikaridai 3–5, Kansai Science City, Kyoto 619–0289, Japan
†† Graduate School of Informatics, Kyoto University　Yoshida-Honmachi, Sakyo-ku, Kyoto, 606–8501 Japan
E-mail: †kuma@nict.go.jp, ††ktanaka@i.kyoto-u.ac.jp

Abstract　We have been developing a retrieval scheme that accurately and exhaustively collects the passages (portion of web pages) which are related to a user-specified topic from the Web. In this article, we propose a method for lexically paraphrasing and expanding user-given queries as the first step of this research. First, our proposed method extracts nouns, adjectives, verbs, and katakana characters as target words from the query or character string which users entered, obtains candidate words for paraphrasing the target words based on information retrieval on the Web, and tests validity of their paraphrasing using two kinds of co-occurrence dictionaries. Then, the method expands the initial query by replacing zero or more of the target words with the candidate words that were determined to be valid. A distinctive point of the method is that it uses not only a co-occurrence dictionary that describes "preceding," "following," and "predicate" relationships between words but also an impression dictionary that describes co-occurrence relationships between words and two contrasting sets of impression words for the validity test. We also verify effectiveness of the method by assessing seven of the paraphrases performed by the method.
Key words　query expansion, opinion mining, question answering, reputation analysis, impression mining

1. ま え が き

Web2.0 時代 [1] の到来に伴い，インターネットはもはや日常的な情報発信手段であり，年齢や性別に関係なく様々な立場の人が様々な情報を発信している．その結果，Web 上には多種多様な情報が大量に存在することとなり，Web 検索システムだけ

でなく，質問応答システム [2] [3] や評判分析システム [4] [5] にとっても貴重な情報源となっている．しかしながら，Web 上の情報の多くは自然言語で記述されており，その記述スタイル（文体や語彙）には個人差があるため，必要な情報を正確かつ網羅的に収集するのは容易でない．そこで著者らは,「育児に参加しない父親」のようなトピック表現（文字列）がクエリとして入力されたときに，そのトピックに関して何らかの記述のある Web 文書を検索し，該当部分（パッセージ）を同定する文字列型パッセージ検索方式を開発することにした．

従来のパッセージ検索方式は，膨大ではあるが限られた数の文書集合を検索対象としており，キーワード集合として与えられるクエリと各文書に含まれる内容語（名詞，形容詞，動詞など）の部分集合との類似度を比較的高コストな手法で計算することにより，パッセージを収集している [6] [7] [8]．そのため，探索空間が限りなく広い Web 検索には向かない．また，クエリにしろ，内容語の部分集合にしろ，単語間の意味的関係が保持されないため，精度悪化の原因となっている．そこで，ユーザが入力したクエリ（文字列）からキーワードを抽出せずに，文字列のまま利用することにする．これにより,（1）通常の Web 検索エンジンを利用して文字列を含む Web 文書を検索できるようになる,（2）単語間の意味的関係が保持され，精度が向上する，といったメリットが生じる．しかしながら，その一方で，検索における制約の厳しさから再現率の低下が懸念されることから，Web 上の情報を正確かつ網羅的に収集するためには，ユーザが入力したクエリをそのまま利用するだけでなく，様々な記述スタイルの正解文書（検索意図に合っていると評価される文書）を収集できるよう，背景にある検索意図に沿ってクエリを展開する必要がある．例えば,「育児に参加しない父親」や「幼児にお年玉をあげる」といったクエリを「子育てに参加しないパパ」や「小さい子どもにお年玉をあげる」に変換する語彙的言い換え [9] [10] [11] や「父親が育児に参加しない」や「幼児にあげたお年玉」に変換する係り受け構造の言い換え [12] [13]，あるいはこれらの組み合わせが必要であり，場合によっては,「子どもが非行に走る」や「幼児の金銭感覚を麻痺させる」に変換するプラン認識ベースの言い換え(注1) も必要になるものと考えられる．

本稿では，以上のような研究の第 1 段階として,「育児」を「子育て」に言い換えるような語彙的言い換え方式を提案する．提案方式は,「怒ってばかりの母親」や「ゴールデンウィークに海外旅行をする」のような名詞句，動詞句，形容詞句をクエリとし,（1）クエリ中の内容語（普通名詞，サ変名詞，形容詞，動詞，カタカナ）に対する言い換え候補を Web 検索により獲得する,（2）言い換えの妥当性を 2 種類の共起辞書を用いて判定する,（3）言い換え可と判定された候補を組み合わせて，新たなクエリ群を生成する，という手順で処理を行う．このとき，単語どうしの前接関係・後接関係・述語関係を示す共起辞書だけでなく，ある特定の印象評価軸に沿って対比させられた 2 つの印象語群との共起関係を示す共起辞書（以降,「印象辞書」と呼び，前述の共起辞書と区別する）を用いて，言い換えの妥当性を判定する点が提案方式の特徴であり，7 個のサンプルクエリを用いた評価実験により，その有効性を検証する．

なお，特定の印象評価軸としては，4 つの印象尺度「期待⇔驚き」,「受容⇔嫌悪」,「喜び⇔悲しみ」,「恐れ⇔怒り」を用意し，それぞれの印象評価軸における値（0〜1 の実数値）を各単語に自動的に付与することによって，印象辞書を構築している．すなわち，印象辞書中の各単語（見出し語）は，この 4 つの値を要素とする印象ベクトルとして表現されている．

以下，2. で関連研究を整理し，提案手法の特徴を明らかにする．3. で言い換えの妥当性を判定する際に用いる事前知識（共起辞書と印象辞書）を新聞記事データベース（約 200 万記事）から自動的に構築する手法について述べ，4. で提案方式の特徴と処理の流れについて述べる．5. で提案方式の有効性を検証し，最後に 6. で本稿のまとめと今後の課題について述べる．

2. 関 連 研 究

2.1 クエリ展開

クエリ展開（query expansion）としては，適合フィードバック（relevance feedback）に基づく方法やシソーラスを用いる方法が一般的である．

適合フィードバックを用いる方法 [14] [15] [16] [17] では,（1）ユーザが入力したクエリを用いて検索を行う,（2）その結果得られた Web 文書集合の一部（例えば，検索結果ランキングの上位 N 件）を検索意図に合っているかどうかという基準でユーザ自身に評価してもらう，あるいはシステムが自動的に評価する,（3）この評価結果に基づいて，クエリに追加すべき検索語を評価された Web 文書の中から抽出したり，クエリから削除されるべき検索語を決定したりする,（4）新たに生成されたクエリを用いて再度検索する，といったことが行われているが，検索結果の中に追加されるべき検索語が含まれていることが前提となっていることから，高い検索精度（precision）を得るためには，最初の検索時に適切なクエリを入力することが要求され，一般ユーザ向きとは言えない．

一方，シソーラスを用いる方法 [18] では,（1）ユーザが入力したクエリ中の検索語の同義語・類義語あるいは関連語をシソーラスから抽出する,（2）検索語を抽出された語と入れ替える,（3）新たなクエリを用いて検索する，といったことが行われており，正解文書の数を増やせることから，再現率（recall）が改善される．しかしながら，その一方で，シソーラスの不備や語の多義性の問題から，検索意図に合わない検索語と入れ替えられ，検索精度の低下を招く場合があることも報告されている．その解決方法として，検索意図に合った検索語が得られるよう，特化したシソーラスを自動構築する手法 [19] [20] や不適切な検索語を共起関係を用いて除去する方法 [21] が提案されており，一定の成果を得ているが，語彙の使い方や文構造といった点で文法的とは限らない Web 文書を文法的であることが前提となっているシソーラスでどの程度まで網羅的に検索できるのかという疑問は残る．

2.2 語彙的言い換え

語彙的言い換えでは，内容語の可換性を記述する大規模な辞書をいかに構築するかが中心的な課題 [22] とされており，WordNet [23] や EDR 日本語単語辞書 [24] のようなシソーラス

(注1)：プラン認識手法を応用すれば，将来的には「父親が遊びに呆けて，家庭を顧みなかったため，子どもが非行に走った」のような文章も検索できるようになるだろう．

を用いる方法 [9] [11]，国語辞典の語釈文を用いる方法 [25] [26]，パラレルコーパスを用いる方法 [27] [28] などが提案されている．

従来の語彙的言い換えでは，機械翻訳や文書要約に資するよう，文法的適格性や文脈的一貫性が重要視されるが，著者らが開発しているパッセージ検索では，(1) Web 文書中に実際に現れる表現であることと (2) 検索意図に合った言い換えであることが重要となる．そこで著者らは，仕様 (1) を実現するために言い換え候補を Web 検索により獲得することにするとともに，仕様 (2) を実現するために以下の設計を導入することにした．すなわち，(a) 単語間の意味的関係（係り受け関係や付属語他によって表現されている内容語どうしの意味的関係）を保持するために，ユーザによって入力された文字列をそのままクエリとするとともに，(b) 言い換えの妥当性を言い換え対象語と候補語の類似性（単語どうしの前接関係・後接関係・述語関係の類似性ならびに対比的な 2 つの印象語群との共起関係の類似性）に基づいて判定することにした．

3. 言い換えに必要な事前知識の自動構築

3.1 共起辞書

単語どうしの前接関係・後接関係・述語関係を表す共起辞書が日経新聞全文記事データベース（1990〜2001 年版）(注2) [29] を解析することにより，自動的に構築される．手順を以下に示す．

まず，記事中の普通名詞，サ変名詞，カタカナを見出し語とし，各見出し語の直前／直後にある名詞（形式名詞と副詞的名詞を除く）やカタカナを前接関係／後接関係として抽出し，直前もしくは直後にある動詞・形容詞（の基本形）を述語関係として抽出する．このとき，動詞や形容詞が名詞を連体修飾している場合を除いて，見出し語との間に 1 個以上の助詞（接続助詞「の」もしくは格助詞）の存在を認めている．また，記事中の動詞や形容詞（の基本形）を見出し語とする場合は，それぞれの直前もしくは直後にある名詞（形式名詞と副詞的名詞を除く）やカタカナを前接関係として抽出する．以上のようにして抽出された共起関係（前接関係，後接関係，述語関係）を見出し語毎にまとめ，それぞれの共起関係を要素，その出現頻度を要素値とするベクトル（共起ベクトル）として共起辞書に登録する．ここで，共起辞書のエントリー例を表 1 に示す．

3.2 印象辞書

印象辞書において，各単語は，4 つの印象尺度「期待⇔驚き」，「受容⇔嫌悪」，「喜び⇔悲しみ」，「恐れ⇔怒り」(注3) に対する印象値（0〜1 の実数値）を要素とする 4 次元の印象ベクトルとして記述される．この印象辞書も日経新聞全文記事データベース（1990〜2001 年版）を解析することにより自動的に構築される．手順を以下に示す．

y 年版に掲載された記事のうち，印象語群 e に含まれる印象語のいずれかを含む記事の数を $df(y,e)$，印象語群 e に含まれる印象語と対象語 w の両方を含む記事の数を $df(y,e\&w)$ とすると，印象語群 e のいずれかが現れたときに，対象語 w も現れる確率 $P(y,e,w)$ は，

$$P = \frac{df(y,e\&w)}{df(y,e)}$$

と表される．次に，対象語 w の印象語群 e_1 に対する出現確率 $P(y,e_1,w)$ と印象語群 e_2 に対する出現確率 $P(y,e_2,w)$ の内分比 $R(y,e_1,e_2,w)$ を

$$R = \frac{P(y,e_1,w)}{P(y,e_1,w)+P(y,e_2,w)}$$

という式で求める．この R 値を各年版毎に求め，平均した結果が対象語 w の印象尺度「$e_1 \Leftrightarrow e_2$」における印象値 $S_{e_1 \Leftrightarrow e_2}(w)$ として印象辞書に登録される．但し，$P(y,e_1,w)+P(y,e_2,w)=0$ となるケースは計算から除外される．

以上の方法で構築された印象辞書の一部を表 2(注4) に示し，登録された単語の数を表 3 にまとめる．また，印象辞書構築の際に，それぞれの印象尺度に対しシードとして与えた印象語群を表 4 に示す．なお，表 2 において，「—」は印象値が設定されていないことを意味する．

表 1　共起辞書のエントリー例（見出し語「記事」の場合）

順位	前接関係		後接関係		述語関係	
1	関連	12,531	社会面	2,390	掲載する	800
2	新聞	789	掲載	318	いう	319
3	特集	326	内容	156	読む	304
4	雑誌	200	スポーツ	141	書く	270
5	インタビュー	191	件数	134	する	237
6	トップ	111	見出し	124	関する	198
7	紹介	111	検索	101	載る	153
8	見出し	94	目	79	載せる	138
9	解説	89	情報	57	つく	132
10	経済	85	データベース	50	出る	122
11	連載	79	写真	50	ある	107
12	紙	67	執筆	40	よる	106
13	企画	64	企業	37	題する	99
14	批判	64	中心	36	見る	98
15	会見	62	一部	33	名誉だ	94
16	死亡	55	コピー	32	紹介する	72
17	誌	47	新聞	32	なる	65
18	問題	45	広告	31	配信する	61
19	掲載	43	切り抜き	30	多い	56
20	観測	41	配信	27	対する	54
	週刊誌	41				
	内容	41				

（数字は共起頻度であり，頻度上位 20 件を表示）

4. 単語印象を考慮した語彙的言い換え方式

図 1 に文字列型のクエリを言い換え，新しいクエリの集合を生成する手順を示す．以下，この手順に従い，著者らが用意した 7 個のサンプルクエリ（表 5）を用いて，提案方式がどのように動作するか，その特徴を示す．

【形態素解析】

汎用の日本語形態素解析システム Juman [31] を用いて，文

(注2)：各年版には，17 万前後の記事（約 200MB）が含まれており，12 年間分で 200 万強の記事が得られた．

(注3)：R. Plutchik が提案した 8 つの基本感情 [30] に基づいている．Plutchik によれば，基本感情は人間が持つすべての感情の基本となる感情（一次感情）であり，他の感情（二次感情）はこの基本感情を混合することで得られる．

(注4)：R 値の計算式が示すように，S 値が 1 に近いほど印象尺度の左側にある語の意味合いが強くなり，0 に近いほど右側にある語の意味合いが強くなる．

表 2　印象辞書のエントリー例

見出し語	印象尺度 期待⇔驚き	受容⇔嫌悪	喜び⇔悲しみ	恐れ⇔怒り
怒る	0.107	0.170	0.274	0.021
母親	0.116	0.179	0.317	0.270
育児	0.285	0.336	0.604	0.404
参加しない	0.468	0.488	0.758	0.602
父親	0.143	0.180	0.298	0.209
ゴールデンウィーク	1.000	1.000	—	—
海外	0.508	0.484	0.638	0.608
旅行	0.309	0.442	0.659	0.425
におい	0.133	0.098	0.485	0.469
きつい	0.397	0.190	0.575	0.422
幼児	0.212	0.297	0.433	0.344
お年玉	0.393	0.516	0.897	0.564
性格	0.299	0.270	0.434	0.395
明るい	0.381	0.298	0.458	0.472
幽霊	0.338	0.416	0.395	0.793
学校	0.207	0.324	0.464	0.346
休む	0.190	0.202	0.554	0.362

表 3　印象辞書に登録された単語の数

	期待⇔驚き	受容⇔嫌悪	喜び⇔悲しみ	恐れ⇔怒り
名詞	86,961	51,677	55,257	49,199
動詞	17,293	13,643	15,173	13,985
形容詞	3,588	3,045	3,473	3,089
カタカナ	30,920	13,368	19,715	11,951

表 4　印象尺度と印象語の対応関係

印象尺度	印象尺度を構成する印象語群
期待⇔驚き	期待（する），予期（する），予想（する），期する ⇔ 驚き，驚く，びっくり（する），驚愕（する），感嘆（する），仰天（する）
受容⇔嫌悪	承知（する），了解（する），了承（する），受け入れ（る） ⇔ 嫌悪（する），嫌う，嫌いだ，嫌だ，毛嫌い（する），忌避（する）
喜び⇔悲しみ	喜び，喜ぶ，うれしい，嬉しい，楽しい，楽しむ，楽しみだ，祝福（する） ⇔ 悲しい，悲しむ，悲しみ，哀しい，哀しみ，悲哀
恐れ⇔怒り	恐れ（る），怖がる，怖い，危惧（する），怯える，恐怖（する） ⇔ 怒り，怒る，憤り，憤る，激怒（する），怒らせる，立腹（する）

字列であるクエリを単語の列に分解する．このとき，単に分解するだけでなく，いくつかの変形操作を行う．例えば,「削除する」は「削除（サ変名詞）」と「する（動詞）」の 2 語からなるが,「削除する（動詞）」に変形し，1 語として扱う．同様に,「楽しくない（形容詞）」や「削除しない（動詞）」も 1 語として扱う．なお，この変形操作は事前知識（共起辞書，印象辞書）の構築時にも行われている．

【Web 検索による言い換え候補獲得】

次に，クエリ中の普通名詞，サ変名詞，形容詞，動詞，カタカナを言い換え対象語とし，クエリから各対象語を取り除いた文字列を言い換え候補獲得のためのクエリとして，Google [32] 上で Web 検索を行う．例えば,「怒ってばかりの母親」というクエリに対しては,「怒る（動詞）」と「母親（普通名詞）」が

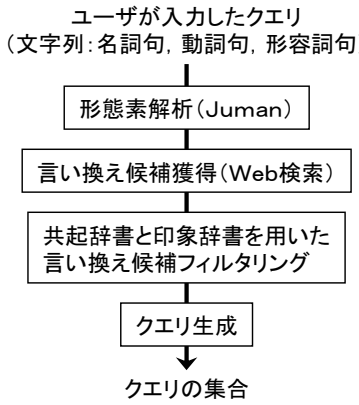


図 1　単語印象を考慮した語彙的言い換え方式

表 5　評価に用いたサンプルクエリ（文字列）

クエリ 1	怒ってばかりの母親
クエリ 2	育児に参加しない父親
クエリ 3	ゴールデンウィークに海外旅行をする
クエリ 4	においのきつい整髪料
クエリ 5	幼児にお年玉をあげる
クエリ 6	性格の明るい幽霊
クエリ 7	海外旅行のために学校を休ませる

言い換え対象語となり,「ばかりの母親」,「怒ってばかりの」という 2 つのクエリが生成され，Web 上で検索される．その結果，対象語のあった場所に来る単語列がそれぞれの対象語に対する言い換えの候補語として扱われる．このとき，対象語の品詞（品詞細分類）に応じて，言い換えの候補語となる単語列の構成が制限される．すなわち，対象語が普通名詞／カタカナのときは候補語は名詞列（1 個以上の名詞（形式名詞，副詞的名詞を除く）・カタカナの連接）であり，サ変名詞のときは名詞列もしくは動詞，形容詞のときは形容詞のみ，動詞のときは動詞もしくはサ変名詞となっている．また，名詞列，動詞，形容詞の前後には 1 個以上の助詞（接続助詞「の」もしくは格助詞）の連接が認められている．

【言い換え候補フィルタリング】

言い換えの妥当性を共起辞書と印象辞書を用いて判定する．

まず，3.1 節で述べた方法で構築された共起辞書を用いて，言い換えの妥当性を判定する．すなわち，共起辞書から言い換え対象語と候補語の共起ベクトルを辞書引きし，次式を用いて類似度（cosine similarity）を求める．この類似度が 0.13 以上のときは，言い換え可，そうでないときは，言い換え不可と判定する．

$$\text{類似度} = \frac{\text{対象語の共起ベクトル} \cdot \text{候補語の共起ベクトル}}{|\text{対象語の共起ベクトル}||\text{候補語の共起ベクトル}|}$$

ここで，表 5 に示した 7 個のサンプルクエリに対するフィルタリングの結果を表 6 に示す．

言い換え可と判定された候補語は，印象辞書を用いて，更にその妥当性を判定される．すなわち，言い換え対象語とその候補語の印象ベクトル間のユークリッド距離を求め，0.36 以下のときは，言い換え可，そうでないときは，言い換え不可と判定する．表 7 にフィルタリングの結果をまとめる．

なお，共起辞書・印象辞書に対する辞書引きは各単語の基本

表 6　共起辞書を用いた言い換え候補フィルタリング

（a）可と判定された言い換え

言い換え対象語	言い換え候補語（類似度）
怒る	叱る（0.19），頷く（0.18），働く（0.18）
母親	お母さん（0.59），男（0.42），自分（0.38），私（0.33），上司（0.24），母（0.24），ママ（0.21），人生（0.19）
育児	子育て（0.16）
参加しない	参加する（0.44），臨む（0.29），対する（0.21），専念する（0.17）
父親	父（0.48），お父さん（0.39），男性（0.34），夫（0.33），子供（0.33），男（0.31），子育て（0.24），状態（0.21），パパ（0.17）
ゴールデンウィーク	旧正月（0.21）
海外	国内（0.45）
する	考える（0.56），予定する（0.22），計画する（0.22），計画（0.19）
におい	匂い（0.82），香り（0.68），ニオイ（0.51），臭い（0.21），香料（0.15）
幼児	子供（0.50），子ども（0.37）
性格	声（0.35），人（0.30）
幽霊	人（0.14）

（b）不可と判定された言い換え

言い換え対象語	言い換え候補語（類似度）
怒る	言い争う（0.08），欲張る（0.05），出産（0.03），喧嘩（0.03），涙（0.02），子育て（0.02），発育（0.01），比較（0.00）
母親	満点（0.12），ストレス（0.09）
育児	家庭（0.06），会話（0.05），活動（0.05），積極（0.04），地域（0.04），行事（0.03），教育（0.03），学校（0.02）
参加しない	かかわる（0.12），適応する（0.03）
父親	具体（0.06），里（0.06）
ゴールデンウィーク	学生（0.01），積極（0.01），シニア（0.01），目的（0.00），月（0.00），流行（0.00）
海外	家族（0.09），ドライブ（0.05），フッ（0.00）
におい	油（0.03）
料	量（0.02）
幼児	一般（0.07），たか（0.07），全員（0.06），全般（0.03），姪（0.02），親兄弟（0.02），親方（0.01），ちん（0.00）
性格	一方（0.07），高校（0.02），台（0.01），青葉（0.00）
幽霊	女性（0.08），タイプ（0.08），子（0.07），犬（0.07），学生（0.06），活動（0.04），茶（0.04），ニーズ（0.02），バー（0.01），トラ（0.00）

形と品詞情報を用いて行われる．そのため,「する」と「される」,「休ませる」と「休んでいた」などは区別されず，常に類似度 1，距離 0 として扱われる．この問題をどう扱うかは今後の課題としたい．また，候補語が名詞列の場合は，その名詞列を構成する各単語と対象語との類似度／距離を算出し，その最大値／最小値を対象語と候補語の類似度／距離とした．それぞれのフィルタリングにおける閾値は実験的に設定された．

【クエリ生成】

表 7　印象辞書を用いた言い換え候補のフィルタリング

（a）可と判定された言い換え

言い換え対象語	言い換え候補語（距離）
怒る	叱る（0.33）
母親	男（0.01），人生（0.02），母（0.08），お母さん（0.08），私（0.12），上司（0.16），自分（0.16），ママ（0.18）
育児	子育て（0.08）
参加しない	参加する（0.21）
父親	父（0.04），男（0.06），お父さん（0.11），夫（0.15），男性（0.19），子供（0.21），パパ（0.25）
海外	国内（0.14）
する	考える（0.17），計画（0.31），予定する（0.32）
におい	臭い（0.19），香り（0.20），匂い（0.30）
幼児	子供（0.08），子ども（0.12）
性格	人（0.10），声（0.18）

（b）不可と判定された言い換え

言い換え対象語	言い換え候補語（距離）
怒る	働く（0.51），頷く（0.72）
参加しない	専念する（0.37），対する（0.39），臨む（0.43）
父親	子育て（0.36），状態（0.56）
ゴールデンウィーク	旧正月（0.81）
する	計画する（0.41）
におい	ニオイ（0.68），香料（0.77）
幽霊	人（0.47）

ユーザが入力したクエリ中の 0 個以上の対象語を候補語と言い換え，新たなクエリとする．例として，表 8 にクエリ 1「怒ってばかりの母親」から生成された全クエリをそのヒット件数とともに示す.「怒って」は「叱って」と言い換えられ(注5),「母親」は「私」,「ママ」,「お母さん」他と言い換えられているのがわかる．なお,「母親」が「男一」という語と言い換えられているが，この候補語は「怒ってばかりの男一匹ガキ大将」という Web 上の文章に由来しており，形態素解析時に行われる変形操作の影響で「男一」だけが名詞列として抽出されている．また,「私」や「自分」といった代名詞も抽出されており，提案方式を有効に活用するためには，検索の結果得られた Web 文書が正解文書であるかどうかを判定することも要求される．このようなツールの開発は今後の課題となる．

5. 考　　察

提案方式の性能を評価し，その有効性について検証する．

【言い換え候補獲得・フィルタリング】

表 5 に示した 7 個のサンプルクエリから，処理「Web 検索による言い換え候補獲得」により全部で 96 語（表 6 参照）(注6) の言い換え候補語が得られた．この 96 語の言い換え妥当性を「常に可（◎），大体の場合において可（○），各クエリの意味（文脈）に

(注5)：言い換え候補の活用形は保持されている

(注6)：「する」と「される」など基本形が同じ場合は，常に言い換え可と判定されるので，この 96 語の中には含まれていない．

表 8 「怒ってばかりの母親」に対するクエリ展開（言い換え）

クエリとなる文字列	ヒット件数
怒ってばかりの私	1,020
怒ってばかりのママ	450
怒ってばかりのお母さん	430
怒ってばかりの自分に	408
叱ってばかりの私	398
叱ってばかりのお母さん	223
怒ってばかりの母親	59
叱ってばかりのママ	38
叱ってばかりの母親	23
怒ってばかりの人生	20
叱ってばかりの自分に	18
怒ってばかりの母が	9
怒ってばかりの上司が	6
怒ってばかりの男一	0
叱ってばかりの母が	0
叱ってばかりの上司が	0
叱ってばかりの人生	0
叱ってばかりの男一	0

（太字はユーザによって入力されたクエリを表す）

おいて可（△），不可（×）」の 4 段階で第一著者が評価したところ，表 9 のような結果が得られた．共起辞書のみを用いた場合のエラーレートは 18.8%（=(3+15)/(41+55)）であったが，印象辞書を併用することにより，12.5%（=(1+2+6+3)/(41+55)）に改善されているのが分かる．共起辞書を用いたフィルタリングにおいて不可と判定された言い換えは，55 語（全候補語の 57.3%）あったが，◎もしくは○と評価されたものはなく，△と評価されたものも，「怒る／言い争う」，「育児／教育」，「怒る／喧嘩」の 3 語だけと好成績であった．また，可と判定された言い換え（41 語）のうち，×と評価されたものは 15 語あったが，うち 9 語は印象辞書を用いたフィルタリングにおいて不可と判定されており，その有効性を示している．残り 6 語は「母親／男」，「母親／上司」，「性格／人」，「性格／声」，「父親／子ども」，「参加しない／参加する」という言い換えであった．パッセージ検索の目的が「トピックに関して何らかの記述のある部分文書を得ること」という意味では，「性格の明るい」を「人の明るい」に言い換えたり，「育児に参加しない」を「育児に参加する」に言い換えたりしても正解と判定されうる文書を収集できるかもしれない．また，表 8 では「母親」が「男一」や「上司が」に言い換えられているが，検索結果は 0〜6 件と少なく，文字列を用いている点が効果的に作用しているのがわかる．一方，印象辞書を用いたフィルタリングにおいて不可と判定された 12 語のうちの 3 語は◎（「におい／ニオイ」）もしくは△（「する／計画する」，「におい／香料」）であり，課題が残る．

【クエリ生成・Web 検索】

クエリ展開によるヒット件数（の総和）の変化を表 10 に示す．表 10 より，大体の場合において，言い換えによるクエリ数の増加に伴い，ヒット件数も増加しており，提案方式の有効性を示しているが，クエリ 7「海外旅行のために学校を休ませる」に対しては，有効に機能していない(注7)．これは，クエリ

表 9　共起辞書と印象辞書を用いた言い換え候補フィルタリング

（a）共起辞書を用いたフィルタリング

	◎	○	△	×	合計
可	11	4	11	15	41
不可	0	0	3	52	55

（b）印象辞書を用いたフィルタリング

	◎	○	△	×	合計
可	10	4	9	6	29
不可	1	0	2	9	12

表 10　クエリ展開によるヒット件数の変化

	展開前	展開後
クエリ 1（18）	59	3,102
クエリ 2（32）	289	739
クエリ 3（16）	5	92
クエリ 4（4）	0	7
クエリ 5（3）	0	90
クエリ 6（3）	1	5
クエリ 7（2）	0	0

（丸括弧内の数字は，生成されたクエリの数を示す）

表 11　クエリ展開による検索精度（precision）の変化

	検索結果	検索精度	
	（Web 文書数）	正解	誤り
展開前	0 (–)	—	—
展開後	7 (2)	5 (100%)	0 (0%)

表 12 「においのきつい整髪料」に対するクエリ展開（言い換え）

クエリとなる文字列	ヒット件数
臭いのきつい整髪料	3
の匂いのきつい整髪料	2
香りのきつい整髪料	2
においのきつい整髪料	0

中の単語数が多かったため，文字列を文脈的制約として利用する提案方式では，獲得できる言い換え候補数が少なくなり，結果，有効なクエリを生成できなかったことが原因と考えられる．このような場合には，単純な語彙的言い換えだけでなく，係り受け構造の言い換えも必要と考えられる．今後の課題としたい．

次に，クエリ展開に伴う検索精度の変化を，クエリ 4「においのきつい整髪料」を用いて調べてみた．このとき，整髪料のにおいに関して何らかの言及がある場合を「正解」，ない場合を「誤り」として計数した．結果を表 11 にまとめるとともに，内訳を表 12 に示す．なお，表 11 において「検索結果」項の括弧内の数字は，すでに削除されていて／書き換えられていて，アクセスできなかった Web 文書の数を示している．

表 11 は，文字列をクエリとすることにより，高精度で正解文書を収集できることを示している．実際，他のクエリに対しても，言い換えが正しく行われているときは，その文脈的制約の厳しさから正解文書が収集されているものと考えられるが，どこまでの言い換えを正しい言い換え（検索意図が合っている）と判定するのかという問題は残る．例えば，「父親」という表現

(注7)：Web 文書中の「海外旅行のために学校を休んでいたそうだ」という文章から「休ませる」の言い換え候補語として「休んでいた」が獲得されているが，「海外旅行のために学校を休んでいた」というクエリに対する検索結果は 0 件となっている．これは Google 検索の特性と考えられる

表 13 「育児に参加しない父親」に対するクエリ展開（言い換え）

クエリとなる文字列	ヒット件数
育児に参加しない父親	286
育児に参加しない男が	241
育児に参加する父親	127
子育てに参加しない父親	18
子育てに参加する父親	18
子育てに参加する男性が	15
育児に参加するパパが	9
育児に参加する男性が	7
育児に参加しない男性が	4
子育てに参加しない夫を	4
育児に参加しない夫を	1
育児に参加しないパパが	1
育児に参加しない里父を	1
育児に参加しないお父さんの	1
育児に参加しないと子供	1
育児に参加する夫を	1
育児に参加する男が	1
子育てに参加しない男が	1
子育てに参加しない男性が	1
子育てに参加するお父さんの	1
育児に参加する里父を	0
育児に参加するお父さんの	0
育児に参加すると子供	0
子育てに参加しないパパが	0
子育てに参加しない里父を	0
子育てに参加しないお父さんの	0
子育てに参加しないと子供	0
子育てに参加する夫を	0
子育てに参加する男が	0
子育てに参加するパパが	0
子育てに参加する里父を	0
子育てに参加すると子供	0

（太字はユーザによって入力されたクエリを表す）

が子供から見た父親のことを言っているのか，それとも自分の父親のことを言っているのか，自明ではないし,「私」や「自分」が母親であるかどうかも定かではない．また,「海外」と「国内」では意味が異なり，従来の言い換えでは不可と判定されるべきものだが,「ゴールデンウィークに海外旅行をする」を「ゴールデンウィークに国内旅行をする」に言い換えても検索意図に合った Web 文書（ゴールデンウィークにおける旅行の大変さを記述する Web 文書）を収集可能であることから，本稿では△（クエリの意味において言い換え可）と評価されている．一方,「参加しない」と「参加する」は正反対の意味であることから，本稿では×（言い換え不可）と評価されているが,「育児に参加する父親」という記述を含む Web 文書は,「育児に参加しない父親」というトピックと裏表の関係と考えられることから，否定語の取り扱いに関しては検討すべき課題と言える．

ここで参考のために，クエリ 2，3，5，6，7 に対して生成されたクエリとそのヒット件数を表 13 から表 17 に示す．なお，表 13 の「里父」は普通名詞の「里」と「父」からなる名詞列であり,「父」と「父親」の類似度・距離を求めることにより，言い換え可と判定されている（表 6，表 7 参照）．

【処理時間】

表 14 「ゴールデンウィークに海外旅行をする」に対するクエリ展開（言い換え）

クエリとなる文字列	ヒット件数
ゴールデンウィークに海外旅行を計画	30
ゴールデンウィークに海外旅行を予定されている	13
ゴールデンウィークに海外旅行を予定されて	13
ゴールデンウィークに海外旅行をする	6
ゴールデンウィークに海外旅行を予定して	6
ゴールデンウィークに国内旅行をする	6
ゴールデンウィークに海外旅行を考えている	5
ゴールデンウィークに海外旅行を予定している	5
ゴールデンウィークに国内旅行を計画	5
ゴールデンウィークに海外旅行をされる	3
ゴールデンウィークに国内旅行を考えている	0
ゴールデンウィークに国内旅行を予定している	0
ゴールデンウィークに国内旅行をされる	0
ゴールデンウィークに国内旅行を予定されている	0
ゴールデンウィークに国内旅行を予定して	0
ゴールデンウィークに国内旅行を予定されて	0

（太字はユーザによって入力されたクエリを表す）

表 15 「幼児にお年玉をあげる」に対するクエリ展開（言い換え）

クエリとなる文字列	ヒット件数
子供にお年玉をあげる	83
の子どもにお年玉をあげる	7
幼児にお年玉をあげる	0

（太字はユーザによって入力されたクエリを表す）

表 16 「性格の明るい幽霊」に対するクエリ展開（言い換え）

クエリとなる文字列	ヒット件数
人の明るい幽霊	3
性格の明るい幽霊	1
の声の明るい幽霊	1

（太字はユーザによって入力されたクエリを表す）

表 17 「海外旅行のために学校を休ませる」に対するクエリ展開（言い換え）

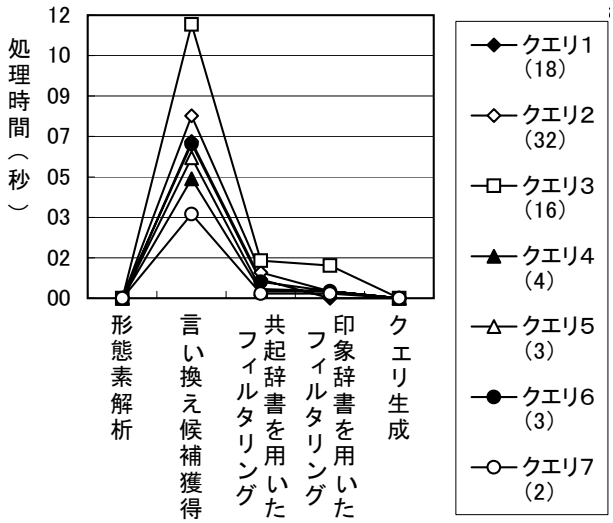
クエリとなる文字列	ヒット件数
海外旅行のために学校を休ませる	0
海外旅行のために学校を休んでいた	0

（太字はユーザによって入力されたクエリを表す）

提案方式が各クエリを処理するのに要した時間を図 2 に示す．言い換え候補獲得を Web 検索に基づいて行っているため，獲得できる言い換え候補数と処理時間はトレードオフの関係になっている．実用化のためにはシソーラスとの併用も考えるべきであろう．

6. ま と め

本稿では,「育児に参加しない父親」のようなトピック表現（文字列）をクエリとするとき，そのトピックと関連のある記述（パッセージ）を Web 上で検索し，収集する文字列型パッセージ検索方式の実現を目指して，語彙的言い換え方式を提案した．提案方式は，何らかのトピック（例えば,「怒ってばかりの母親」）をクエリとし,（1）クエリ中の内容語（普通名詞，サ変名詞，形容詞，動詞，カタカナ）に対する言い換え候補を Web



（丸括弧内の数字はクエリ数を示す）

図 2　各処理に要する処理時間

検索により獲得する，(2) 言い換えの妥当性を共起辞書と印象辞書を用いて判定する，(3) 最終的に言い換え可と判定された候補語を用いて，新たなクエリ群を生成する，という手順で処理を行う．このとき，単語どうしの前接関係・後接関係・述語関係を示す共起辞書だけでなく，ある特定の印象評価軸に沿って対比させられた 2 つの印象語群との共起関係を示す印象辞書を用いて，言い換えの妥当性を判定する点に特徴があり，7 個のサンプルクエリを用いた評価実験により，その有効性を検証した．

今後の課題としては，大規模な被験者実験に基づく提案手法の定性的・定量的評価，検索結果が検索意図に合っているかどうかを判定するためのツールの開発，他の言い換え方式（多対多の語彙的言い換え，係り受け構造の言い換えなど）の開発，等が挙げられる．

文　献

[1] 大向一輝，橋本大也（編著）：特集「Web2.0 の現在と展望」，情報処理学会誌，Vol.47，No.11，pp.1193–1236 (2006).

[2] Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal: Probabilistic Question Answering on the Web, Proc. of the International WWW Conference, Honolulu, USA (2002).

[3] 藤井敦：百科事典としての WWW，人工知能学会誌，Vol.19，No.3，pp.296–301 (2004).

[4] 立石健二，石黒義英，福島俊一：インターネットからの評判情報検索，情処研報，自然言語処理 144-11，pp.75–82 (2001).

[5] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima: Mining Product Reputations on the Web, Proc. of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp.341–349, Edmonton, Alberta, Canada (2002).

[6] Gerard Salton, James Allan, and Chris Buckley: Approaches to Passage Retrieval in Full Text Information Systems, Proc. of the ACM SIGIR Conference on R&D in Information Retrieval, pp.49–58 (1993).

[7] James P. Callan: Passage-Level Evidence in Document Retrieval, Proc. of the ACM SIGIR Conference on R&D in Information Retrieval, Dublin, Ireland, pp.302–310 (1994).

[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，自然言語処理，Vol.6，No.3，pp.101–126 (1999).

[9] Ingrid Zukerman, Sarah George, and Yingying Wen: Lexical Paraphrasing for Document Retrieval and Node Identification, Proc. of the 2nd International Workshop on Paraphrasing, Vol.16, Sapporo, Japan, pp.94–101 (2003).

[10] Takenobu Tokunaga, Hozumi Tanaka, and Kenji Kimura: Paraphrasing Japanese Noun Phrases Using Character-based Indexing, Proc. of the 2nd International Workshop on Paraphrasing, Vol.16, Sapporo, Japan, pp.80–87 (2003).

[11] 近藤恵子，佐藤理史，奥村学：「サ変名詞＋する」から動詞相当句への言い換え，情報処理学会論文誌，Vol.40，No.11，pp.4064–4074 (1999).

[12] 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，情処学自然言語処理研報，Vol.2000，No.011，pp.181-188 (2000).

[13] 竹内淳平，辻井潤一：係り受け関係と言い換え関係を用いた柔軟な日本語検索，言語処理学会第 11 回年次大会発表論文集，pp.568–571 (2005).

[14] Joseph John Rocchio: Relevance Feedback in Information Retrieval, In The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs, New Jersey (1971).

[15] Gerard Salton: Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing Co., Inc. Boston, USA (1989).

[16] 新田清，蓬莱尚幸，薗部正幸：文書クラスタリングを利用した検索質問展開手法の開発と評価，情処研報，データベースシステム 118-2，情報学基礎 54-2，pp.9–16 (1999).

[17] 帆足啓一郎，松本一則，井ノ上直己，橋本和夫：文書間の類似度における単語寄与度を利用した検索式拡張手法，情報処理学会論文誌：データベース，Vol.40，No.SIG 8 (TOD 4)，pp.63-73 (1999).

[18] Ellen M. Voorhees: Query Expansion using Lexical-Semantic Relations, Proc. of the 17th ACM SIGIR Conference on R&D in Information Retrieval, pp.61–69 (1994).

[19] 佐々木稔，新納浩幸：潜在的文脈関連度を用いた検索質問拡張，情処研報，自然言語処理 151-10，情報学基礎 68-10，pp.65–72 (2002).

[20] 高野敦子，平井誠：新聞記事検索における観点を考慮したクエリー拡張手法，情処研報，自然言語処理 152-16，pp.107–114 (2002).

[21] 栗山和子：シソーラスを用いた検索式拡張の評価，情処研報，情報学基礎 52-1，pp.1–8 (1998).

[22] 乾健太郎：言語表現を言い換える技術，言語処理学会第 8 回年次大会チュートリアル資料集，pp.1–21 (2002).

[23] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller: Introduction to WordNet: An On-line Lexical Database, International Journal of Lexicography, Vol. 3, No. 4, pp.235–312 (1990).

[24] EDR HomePage: `http://www2.nict.go.jp/r/r312/EDR/J_index.html`.

[25] 藤田篤，乾健太郎：語釈文を利用した普通名詞の同概念語への言い換え，言語処理学会第 7 回年次大会講演論文集，B5-1，pp.331–334 (2001).

[26] 鍛治伸裕，黒橋禎夫，佐藤理史：国語辞典に基づく平易文へのパラフレーズ，情報処理学会研究報告，2001-NL-144，pp.167–174 (2001).

[27] 今村賢治，秋葉泰弘，隅田英一郎：階層的句アライメントを用いた日本語翻訳文の換言，第 7 回言語処理学会年次大会ワークショップ，pp.15–20 (2001).

[28] 関根聡史：複数の新聞を使用した言い替え表現の自動抽出，第 7 回言語処理学会年次大会ワークショップ，pp.9–14 (2001).

[29] 日経全文記事データベース DVD-ROM 版，1990–1995 年版，1996–2000 年版，2001 年版，日本経済新聞社.

[30] Robert Plutchik: The Emotions: Facts, Theories, and a New Model, New York: Random House (1962).

[31] 黒橋禎夫，河原大輔：日本語形態素解析システム JUMAN version 4.0 (2003).

[32] Google, `http://www.google.co.jp/`