D-049

2種類の共起情報を用いた語彙的言い換えに基づくWeb検索

Web Retrieval Based on Lexical Paraphrasing Using Two Kinds of Cooccurrence

熊本 忠彦 †
Tadahiko Kumamoto

田中 克己 ††
Katsumi Tanaka

1. まえがき

Web2.0時代と呼ばれる昨今，Web上には多種多様な情報が大量に存在しているが，その記述スタイル（特に文体や語彙）には個人差があるため，必要な情報を幅広く網羅的に収集するのは容易でない．そこで本論文では，ユーザが入力したクエリ（名詞句，動詞句，形容詞句）を語彙レベルで言い換え，新たなクエリ群を生成することによって，より多くの正解文書（検索意図に合っていると評価される文書）を収集可能にするWeb検索方式を提案する．

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」のようなトピック表現（文字列）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語を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語を組み合わせて，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と特定の印象評価軸において対比的な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）の2種類を用いて，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．

2. 設計コンセプト

従来の語彙的言い換えでは，シソーラスを用いる方法[1][2]や国語辞典の語釈文を用いる方法[3]，パラレルコーパスを用いる方法[4][5]などが提案されているが，機械翻訳や文書要約，文生成への応用を目的としているため，文法的適格性や文脈的一貫性が重要視されており，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[6]となっている．

一方，Web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の2点が重要と言える．提案方式では，この2つの要件を満たすために以下のような設計コンセプトを導入している．(a) 言い換え候補語をWeb検索により獲得する．これにより，Web文書中で実際に使われている文表現からの候補語抽出が可能となる．(b) 言い換えの妥当性を言い換え対象語と候補語の「使われ方」に関する類似性（単語どうしの前接関係・後接関係・述語関係の類似性）と「心的印象」に関する類似性（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．これにより，使われる場面・状況に応じた言い換えが可能となる．

なお，提案方式では，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Web検索エンジンに投入している．その結果，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）が保持され，より精度の高いWeb検索が可能となっている．

表1: 共起辞書の一部（見出し語「記事」の場合）

前接関係		後接関係		述語関係	
関連	12,531	社会面	2,390	掲載する	800
新聞	789	掲載	318	いう	319
特集	326	内容	156	読む	304
雑誌	200	スポーツ	141	書く	270
インタビュー	191	件数	134	する	237
トップ	111	見出し	124	関する	198
紹介	111	検索	101	載る	153
見出し	94	目	79	載せる	138
解説	89	情報	57	つく	132
経済	85	データベース	50	出る	122

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を日経新聞全文記事データベース（1990～2001年版）*[7]を解析することにより，構築した．

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．一方，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）が見出し語毎に整理され，共起関係を要素，その共起頻度を要素値とする共起ベクトルとして共起辞書に登録された．共起辞書の一部（各関係における頻度上位10件）を表1に示す．

3.2 印象辞書

印象辞書も共起辞書と同様，日経新聞全文記事データベース（1990～2001年版）を解析し，構築した．

まずはじめに，Plutchikが提案した8つの基本感情[8]に基づいて，4つの印象評価軸「期待⇔驚き」，「受容⇔嫌悪」，「喜び⇔悲しみ」，「恐れ⇔怒り」†を設定し，

† 千葉工業大学・情報科学部・情報ネットワーク学科
†† 京都大学大学院・情報学研究科・社会情報学専攻

*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．

†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．

表 2: 印象評価軸と印象語の対応関係

軸	印象評価軸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， 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）

表 3: 印象辞書の一部

見出し語	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
父親	0.143	0.180	0.298	0.209
におい	0.133	0.098	0.485	0.469

表 4: 印象辞書に登録された単語の数

品詞	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

それぞれの軸において対比的な 2 つの印象語群（表 2 参照）をシソーラスを参考に定義した．

次に，y 年版に掲載された記事のうち，印象語群 e に含まれる印象語を 1 語以上含む記事の数を $df(y,e)$，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，

$$P = df(y,e\&w)/df(y,e)$$

と表される．したがって，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 $R(y,e_1,e_2,w)$ は，

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

という式で求められる．この R 値を各年版毎に求め，平均した結果が対象語 w の印象評価軸「$e_1 \Leftrightarrow e_2$」における値 $S_{e_1 \Leftrightarrow e_2}(w)$ となる．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た．

印象評価軸は 4 つあるので，印象辞書中の各単語（見出し語）には，各軸において算出された 4 つの値を要素とする印象ベクトルが登録された．ここで印象辞書の一部を表 3[‡] に示し，登録された単語の数を表 4 にまとめる．

4. 語彙的言い換えに基づく Web 検索方式

図 1 に提案方式の処理手順を示す．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 参照）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．

[‡] R 値の計算式が示すように，S 値が 1 に近いほど軸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．

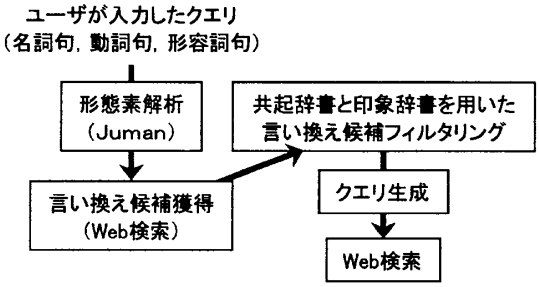


図 1: 語彙的言い換えに基づく Web 検索方式

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

【形態素解析】

日本語形態素解析システム Juman[9] を用いて，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）」のような否定語も 1 語として扱う．なお，以上の変形操作は共起辞書・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[10] 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】

まず，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，コサイン類似度を求める．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．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

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

対象語	候補語（コサイン類似度）
怒る	叱る (0.19)，頷く (0.18)，働く (0.18)
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)
におい	匂い (0.82)，香り (0.68)，ニオイ (0.51)，臭い (0.21)，香料 (0.15)

(b) 不可と判定された言い換え

対象語	候補語（コサイン類似度）
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)
父親	具体 (0.06)，里 (0.06)
におい	油 (0.03)

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

対象語	候補語（ユークリッド距離）
怒る	叱る (0.33)
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)
におい	臭い (0.19)，香り (0.20)，匂い (0.30)

(b) 不可と判定された言い換え

対象語	候補語（ユークリッド距離）
怒る	働く (0.51)，頷く (0.72)
父親	子育て (0.36)，状態 (0.56)
におい	ニオイ (0.68)，香料 (0.77)

表 8: 「怒ってばかりの母親」に対する言い換え

生成されたクエリ（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

（太字はユーザによって入力されたクエリを表す）

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示す．

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】

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 から生成されたクエリの一部（ヒット件数上位 10 件）を示す．「怒って」は「叱って」と言い換えられ[§]，「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，「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，照応解析といった言語処理技術の導入が必要と考えられる．このような技術の導入・開発は今後の課題とする．

5. 考察

提案方式の性能を評価し，その有効性について考察する．

【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全部で 96 語（表 6 参照）[¶] 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誤り率は 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．なお，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，× 以外であった 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】

クエリ生成によるヒット件数（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[‖]．これは，クエリ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，クエリ生成に伴う検索精度の変化を調べてみた．その結果，言い換えが正しく行われているときは，

§ 言い換え候補語の活用形は保持されている．

¶ 「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．

‖ 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 9: 言い換え候補フィルタリングの妥当性評価

(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

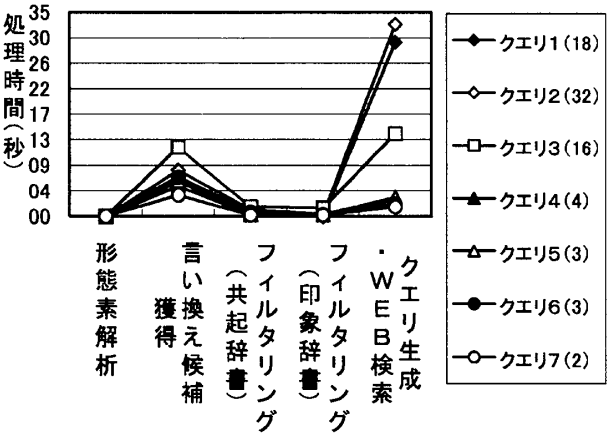
(b) 印象辞書を用いたフィルタリング

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

表 10: クエリ生成によるヒット件数の変化

	生成前	生成後
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）



（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

高精度で正解文書を収集できることがわかった．これはトピックそのものである文字列をクエリとしていることから当然の結果と言えるが，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．

【処理時間】

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）がクエリとして入力されたときに，その内容語を言い換えることによって新たなクエリ群を生成し，より多くの正解文書を検索可能にする Web 検索方式を提案した．提案方式は，言い換えの妥当性を判定するために，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，特定の印象評価軸において対比的な 2 つの印象語群との共起関係を示す印象辞書を用いている点に特徴があり，7 個のサンプルクエリを用いた評価実験により，その有効性が示された．

今後の課題としては，大規模な被験者実験に基づく提案方式の定性的・定量的評価，代名詞等の照応解析を行うツールの開発，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．

参考文献

[1] 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).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 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.

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

[9] 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).

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