社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－NL－176（4）
2006／11／22

同義語を用いた質問文の拡張による係り受け関係の柔軟な照合

西村　涼　渡辺　靖彦　岡田　至弘

龍谷大学 理工学部 情報メディア学科
〒520-2194 大津市瀬田大江町横谷 1-5

E-mail: r_nishimura@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

あらまし　質問文に現れる用言の同義語を検索対象のドキュメントから自動抽出し、その中から適切な語をユーザに選択させることで、質問文の拡張を行い、係り受け関係の柔軟な照合に利用する。

キーワード　同義語、拡張検索、係り受け関係の柔軟な照合

Flexible Matching of User Query Expanded Using Similar Predicates

Ryo NISHIMURA, Yasuhiko WATANABE, and Yoshihiro OKADA

Dept. of Media Informatics, Ryukoku University
Seta, Otsu, Shiga, 520-2194, Japan

E-mail: r_nishimura@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp

Abstract　In this paper, we propose a method of expanding user's question by using similar predicates. When a user gives a question to our system, it extracts some predicates from targeted documents, which seem to be similar predicates, and shows them to the user. Then, similar predicates which are selected by the user are used for expanding user's question and flexible matching of dependency structure.

Key words　similar predicate, query expansion, flexible matching of dependency structure

1. はじめに

自然な文による情報検索ではさまざまな表現を用いることができるので、同義表現のあつかいが重要である。同義表現にはさまざまな種類があり、用言の同義語に限定しても、とりあつかいがむずかしい例が多い。以下の例は、vine linux に関心のある人たちが情報を交換しているメーリングリスト (Vine Users ML (注1)) に投稿された問合わせの文である。

(例 1a) apache で CGI が使用できない

(例 1b) apache で CGI が実行できない

「使用する」ことと「実行する」ことは一般に同義ではない。しかし、この場面 (コンピュータを働かせる場面) では、これらの用言は同じ意味を表わしている。この例のような同義語関係は、シソーラスや辞書の語義説明文などには記述されていないので、取り扱うのはむずかしい。しかし、こうした同義語関係を取り扱えないと、ユーザが用いなかった表現で記述されている情報を検索するのに失敗してしまう。次の例も Vine Users ML に投稿された問合わせ文であるが、(例 1) の場合以上に取り扱いがむずかしい例である。

(例 2a) IP アドレスが割り当てられません

(例 2b) IP アドレスが取得できません

(例 1) では、場面や状況、対象物 (apache と CGI) に対する認識に違いはなかった。一方、(例 2) では、対象物 (IP アドレス) に対する認識が異なっている。(例 2a) の発話者は「IP アドレスは与えられるものである」と考えている。一方、(例 2b) の発話者は、「IP アドレスは手に入れるものである」と考えている。このように、対象物 (IP アドレス) に対する認識が異なっているため、同じ内容を表わしているにもかかわらず、一般に同義語ではない用言「取得する」と「割り当てる」が用いられている。このような場合についても同義語と判定できることがのぞましい。

そこで、用言がとる格要素を比較して、検索対象の文で用いられる用言の中からユーザが質問文で用いた用言の同義語である可能性が高いものを取り出す方法を提案する。本研究では、検索対象の文としてメーリングリストに投稿された問合わせ文を用いる。提案する方法によって、シソーラスや国語辞典に記述されているような一般に成り立つ同義語関係だけではなく、

- 特定の場面や状況下で生じる同義語関係
- 場面・状況・対象物への認識のずれによって生じる同義語関係

を取り出せることを示す。さらに、提案する方法がユーザの質問文を拡張するのに利用できることを示す。

2. 関連研究

自然な文で検索要求を表現させるシステムは多く、異表記同義語の問題もユーザの質問文を拡張するためにさかんに研究されている。シソーラスを用いてユーザの質問を拡張すると、再現率は向上するが、精度は下がることが報告されている [1]。しかし、奥村ら [2] は、単語の表記に基づく辞書以外の概念間の関係を表わす辞書を用いることで、精度が下がらないことを示した。

ユーザの質問文を拡張するには、シソーラスなどが用いられることが多い。シソーラスを人手ではなく、自動構築する研究もさかんに行われている。上野ら [3] は、係り受けの 2 部グラフと文脈情報を用いて大量の文書から同義・類義表現の候補を取り出し、辞書の作成に利用することを提案している。鍛冶ら [4] は、国語辞典で説明されている同義・類義表現を利用できる場合にはかなりよい精度でいいかえができることを示した。しかし、(例 1) の場合のように限定

(注1)：http://vinelinux.org/ml.html

された場面や状況下で生じる同義語関係や、(例 2) の場合のように認識のずれによって生じた同義語関係などは辞書やシソーラスに記述されていないのであつかえない。

3. 用言の同義語候補の抽出

3.1 用言のさまざまな同義語関係

用言の同義語関係にはさまざまなものがある。ユーザの質問文を拡張するのに重要であると考えたものを以下に示す。

最初の例は、特定の場面や状況だけではなく、一般に成り立つ同義語関係の例である。

(例 3a) 音が鳴る

(例 3b) 音が出る

大辞林 (第二版) の語義説明文では「鳴る」は「音が出る」と説明されている。この例のような場合では、シソーラスや語義説明文を利用してユーザの質問文を拡張できるので、ユーザが用いなかった用言で記述されている情報も検索できる。

一方、1. 章で示した (例 1) は特定の場面や状況で生じる同義語関係の例であった。

(例 1a) apache で CGI が使用できない

(例 1b) apache で CGI が実行できない

「使用する」ことと「実行する」ことは一般に同義ではない。しかし、この場面 (コンピュータを働かせる場面) では、これらの用言は同じ意味を表わしている。この例のような同義語関係は、シソーラスや辞書の語義説明文などには記述されていないので、取り扱うのはむずかしい。しかし、こうした同義語関係を取り扱えないと、ユーザが用いなかった用言で記述されている情報を検索するのに失敗してしまう。

(例 1) と (例 3) で同義関係にある用言が表現する意味はほぼ同じであった。一方、以下の例の用言はいいかえが可能であるが、表現する意味は異なる。

(例 4a) MO を使用できない

(例 4b) MO を接続できない

「使用する」は「接続する」といいかえることはできるが、それぞれが表現している内容は異なる。「接続する」は「使用する」場合の 1 つで、このほかにも

- マウントする
- アクセスする
- アンマウントする

などが「使用する」といいかえることができる。これらの表現は「使用する」とは厳密には同義語関係ではないが、ユーザの質問文を拡張するためには、その取り扱いが重要になる例である。

さらに、1. 章の (例 2) のように、場面や状況、対象物に対する認識のずれによって生じる同義語関係もある。

(例 2a) IP アドレスが割り当てられません

(例 2b) IP アドレスが取得できません

この例では、対象物 (IP アドレス) に対する認識が (例 2a) と (例 2b) では異なっている。(例 2a) の発話者は「IP アドレスは与えられるものである」と考えている。一方、(例 2b) の発話者は「IP アドレスは手に入れるものである」と考えている。このように、対象物 (IP アドレス) に対する認識が異なっているため、同じ内容を表わしているにもかかわらず、一般に同義語ではない用言「取得する」と「割り当てる」が用いられている。このような場合についても同義語と判定できることがのぞましい。

3.2 用言の同義語抽出の処理の概要

ユーザが入力した質問文中の用言と同義語である可能性が高い用言を、検索対象の文から取り出す処理の概要を図 1 に示す。検索対象の文には、メーリングリストに投稿された問合わせのメールを利用した。メーリングリストに投稿された問合わせメールを利用するのは、以下に示す有利さがあるからである。

- メーリングリストでは特定の目的に

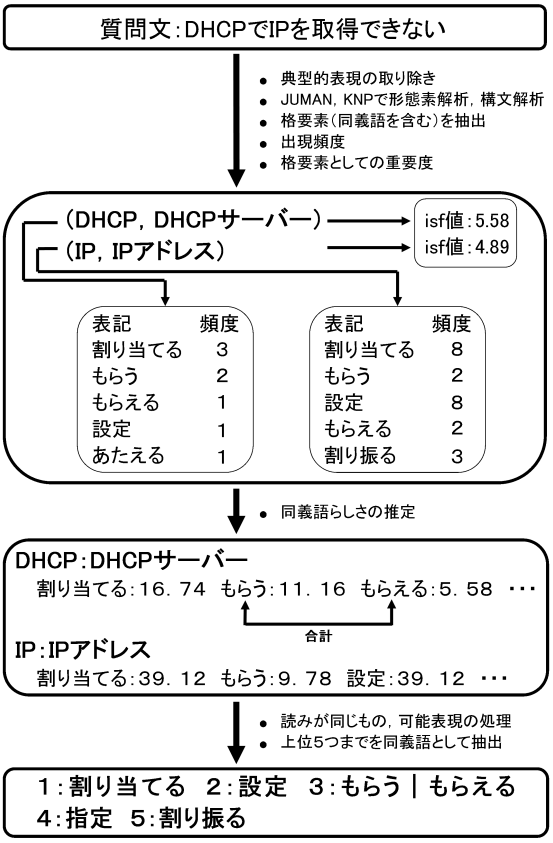

図 1　同義語抽出の処理の概要

ついて情報交換を行っているので、特定の場面や状況下での問合わせ文を収集することができる。

- さまざまなメーリングリストがあるので、さまざまな場面や状況下での問合わせ文を収集することができる。

ただし、問合わせメールには問合わせそのものとは関係のない文も含まれている。そこで、渡辺ら [5] の方法で問合わせメールから問合わせの中心になる文 (重要文) とその前後 1 文を取り出し、それらの文に含まれる用言からユーザが入力した質問文中の用言と同義語である可能性が高いものを取り出すことにした。

本研究では、格要素を比較することによって、ユーザの質問文中の用言と同義語関係にある可能性が高い用言を検索対象の文から取り出す方法を提案する。これは、同義語関係にある用言はよく似た格要素をもつことがあるからである。格要素になる体言 (名詞およびサ変名詞) は用言に比べて、

- 特定の場面や状況下で生じる同義語関係は少ない
- 認識のずれが原因で生じる同義語関係も少ない

このため、特定の場面や状況下でも、あるいは認識のずれがあっても、体言の同義語関係の取り扱いは用言ほどむずかしくない。これは、格要素となる体言の同義・類義関係を手がかりにして

- 特定の場面や状況下で生じる用言の同義語関係
- 認識のずれが原因で生じる用言の同義語関係

など用言のさまざまな同義語関係を取り出そうとするのには有利である。

そこで、ユーザの質問中にふくまれる用言 V_{user}(N 個の格要素 C_i をもつ) と検索対象の文に含まれる用言 V_{target} の類似度、すなわち同義語らしさ $Sim(V_{user}, V_{target})$ を以下のように定義した。

$$Sim(V_{user}, V_{target}) = \sum_{i=1}^{N} f_{dp}(C_i, V_{target}) \cdot isf(C_i)$$

ただし、$f_{dp}(C_i, V_{target})$ は、体言 C_i およびその同義語を格要素としてもつ V_{target} の数である。$isf(C_i)$ は体言 C_i の格要素としての重要度をあらわす値で、以下の式で表わす。

$$isf(C_i) = log\left(\frac{N}{sf(C_i)}\right)$$

N は検索対象の文の総数で、その中で体言 C_i を含む文の数が $sf(C_i)$ である。

3.3　用言の同義語らしさの推定手順

検索対象の文で用いられている用言から、ユーザの質問文で用いられている用言と同義語である可能性が高いものを取り出す手順を以下に示す。検索対象の文には、メーリングリストに投稿された問合わせメールから取り出した文を用いた。

step 1 Vine Users ML に投稿された問合わせメールから、問合わせの中心になる文とその前後の 1 文を渡辺らの方法 [5] で取り出す。取り出した文は形態素解析と係り受け解析を行う。形態素解析には JUMAN [6]、係り受け解析には KNP [7] を用いた。

step 2 ユーザが入力した質問文から、質問の内容にかかわらない典型的な表現をとりのぞく。具体的には、

- ～について教えてください
- ～について知りたい
- ～方法を教えてください
- ～方法がわかりません
- ～どうしたらいいのでしょうか

などの表現が含まれていたら、それらの表現をとりのぞく。

step 3 step 2 の結果に対して形態素解析と係り受け解析を行う。形態素解析には JUMAN [6]、係り受け解析には KNP [7] を用いた。係り受け解析の結果から用言とその格要素の組を取り出す。

step 4 ユーザの質問文で用いられている用言 V_{user} とメーリングリストに投稿された問合わせメールから取り出した文で用いられている用言 $V_{ml}(j)$ $(j = 1, 2, \cdots, M)$ の類似度 $Sim(V_{user}, V_{ml}(j))$ を 3.2 節で述べた方法で計算する。

step 5 step 4 の結果では、読みが同じ用言 (例:「割り当てる」と「わりあてる」) や可能表現 (例:「もらう」と「もらえる」) がそれぞれ別の用言として扱われている。そこで、それらの用言の step 4 の結果を合計し、1 つの用言として取り扱う。

step 6 step 5 の結果から、V_{user} と類似度の高い用言を 5 つまで取り出す。

4. 実験結果と評価

4.1 同義語候補の抽出

3. 章で述べた手法により、同義語候補の抽出を行い、その結果を評価した。実験では検索対象の文として、Vine Users ML に投稿された 8782 通の問合わせのメールから取り出した

- 質問の中心になる文の前の文 (7330 文)
- 質問の中心になる文 (8782 文)
- 質問の中心になる文の後の文 (8614 文)

の合計 24726 文を用いた。ユーザの質問文として用いた 32 個の質問を図 2 に示す。これらの質問文は、Vine Users ML に類似した内容の情報交換を行っているメーリングリスト Linux Users ML に投稿された質問から取り出した。32 個の質問文に含まれていた 35 個の用言の同義語の候補を検索対象の文から 162 個取り出した。取り出した同義語の候補を評価し、以下の 3 種類に分類した。

○ 取り出した用言とユーザの質問文だけをみて、その用言が同義語であると判定できるもの

△ 用言を取り出した文を読むと同義語であると判定できるが、その用言とユーザの質問文だけをみても同義語かどうか判定するのがむずかしいもの

× 同義語ではないと判定できるもの

抽出できた同義語候補の判定結果を表 1 に示す。実験に用いた 32 個の質問のうち同義語の候補が抽出できた質問は 28 個であった。同義語候補が抽出できなかった 4 個の質問 (質問 3、7、11、22) は、質問文中の用言がもつ格要素と同じものをもつ用言を検索対象文ではみつけられなかった例である。

○ と評価された用言の例を図 3 に示す。○ と評価された用言の中には、「取得」と

(1) DHCP で IP を再取得できない
(2) Linux で音が出ません
(3) XWindowSystem 起動時の不都合について
(4) ハードディスクのパーティションの修復
(5) Apache に SSI を許可する設定はいずこに
(6) proftpd にログインできない
(7) 漢字入力できません
(8) NIC を二枚使用して、Linux マシーンをルータとして機能する方法を教えてください
(9) Apache1.39 で CGI が使えない
(10) 再起動すると時間がくるう
(11) 英語エラーメッセージに戻す方法がありましたらお教え下さい
(12) NFS サーバが起動しません
(13) MO を使う方法を教えてください
(14) トラフィックのモニタリングする方法はありませんでしょうか
(15) Emacs で漢字コードを指定するにはどうしたらいいのでしょうか
(16) X で＼キーが入力できない
(17) PDF のテキストだけを抽出する方法を教えてください
(18) login するときに時間がかかってしまいます
(19) lpr で印刷ができないで困っています
(20) Emacs でバックアップファイルをつくらない方法を教えてもらえないでしょうか
(21) Xwindow の画面を取り込むにはどうしたらいいのでしょうか
(22) レスキューディスクがないときの起動はできるのでしょうか
(23) PCMCIA スロットを使えるように設定したのですが、ネットワークカードをネットワークカードとして認識してくれません。
(24) PPxP が実行できない
(25) chmod ができる FTP サーバを探しています
(26) Makefile の記述方法がわかりません
(27) 特定のユーザを telnet でログインできないようにしたいのですが、どういう設定が必要なのか教えていただけないでしょうか?
(28) VineLinux2.5 で Webmin を起動しようとすると、localhost:10000 へのネットワーク接続を試みているときに接続が拒否されました
(29) 自作マシンにビデオキャプチャーカードを挿したはいいもの xawtv を用いてテレビを見ることができません
(30) LaTex で書かれた日本語の文章があって、これを Word の文章にしたい
(31) リソースを監視できるソフトの中でお勧めのソフトって何かありますでしょうか
(32) CDROM の mount ができずにてこずっています

図 2　Linux Users ML から取り出した 32 個の質問

表 1　同義語候補の判定結果

評価	○	△	×	Total
用言の数	39	39	84	162

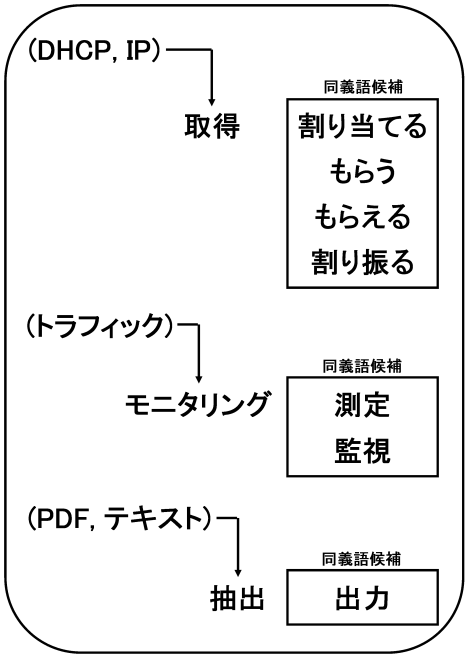

図 3　○ と評価された用言の例

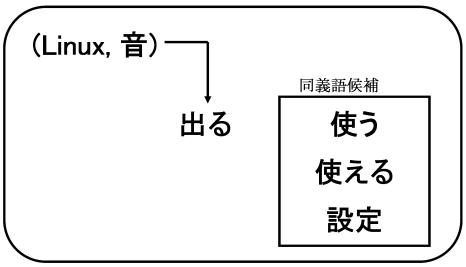

図 4　△ と評価された用言の例

「割り当てる」のように、一般に同義関係にはないのだが、認識のずれのため同義語関係が生じたものが含まれている。

△ と評価された用言の例を図 4 に示す。△ と評価された用言の中には、「使う」「使える」「設定」など「音が出る」の「出る」と同義語かどうか判定がむずかしい用言が含まれている。「設定」は「音の設定」などから取り出されたもので、取り出された文をみるとユーザの用いた用言と同義語であると判定できるものである。

表 1 に示すように、同義語抽出の結果か

ら同義語辞書を作成するのはむずかしい。しかし、ユーザの質問文に対し、同義語関係がある用言の候補を示し、その中から同義語として適切なものを選ばせ、ユーザの質問文を対話的に拡張するのには本手法は有効であると考えられる。同義語を○と△に分類したのは、用言だけを示してユーザに同義語かどうか簡単に判定できるものが取り出した用言の中にどれだけ含まれているのかを調べるためである。次節では、取り出した同義語の候補を利用することで、ユーザの質問文を拡張できることを示す。

4.2 同義語を用いた質問文の拡張による係り受け関係の柔軟な照合の実験と評価

4.1節で取り出した同義語の候補を利用してユーザの質問文を拡張し、その係り受け関係の照合結果について調べた。ユーザの質問文として図2の32個の質問のうち、同義語候補を抽出できなかった4個の質問をのぞいた28個の質問を用いた。検索対象の文にはVine Users MLに投稿された8782通の問合わせのメールから取り出した

- 質問の中心になる文の前の文
- 質問の中心になる文
- 質問の中心になる文の後の文

の合計24726文を用いた。照合結果は、文の構文的な構造と単語の重要度にもとづいて順位づけた[5]。そして、それぞれの質問に対する照合結果の上位10位、20位、30位のものまで調べ、以下の4種類の方法で評価した。表2に評価結果を示す。

評価1 質問を拡張しない場合、調査した照合結果の中に係り受け関係が照合されたものがどれだけあるか

評価2 ○と評価された同義語で質問を拡張した場合、調査した照合結果の中に係り受け関係が照合されたものがどれだけあるか

評価3 ○と△と評価された同義語で質問を拡張した場合、調査した照合結果の中に係り受け関係が照合されたものがどれだけあるか

評価4 ○と△と×と評価された同義語、すなわちすべての同義語候補で質問を拡張した場合、調査した照合結果の中に係り受け関係が照合されたものがどれだけあるか

表2 同義語を用いた質問文の拡張による係り受け関係の照合結果

評価方法	評価1	評価2	評価3	評価4
上位10個	59	119	158	221
上位20個	93	188	257	368
上位30個	137	251	344	509

拡張なしでの照合結果である評価1に比べて、○と評価された用言を用いて質問文を拡張した評価2では係り受け関係の照合結果がおよそ2倍になっている。評価2で質問を拡張するのに用いた用言は、ユーザの質問文をみるだけで同義語として適切だと簡単に判定できるものである。したがって、提案手法で取り出した用言は、ユーザの質問を拡張して係り受け関係を照合するのに有効であるといえる。

次章で述べる検索システムでは、ユーザは用言だけをみて、自分が用いた用言と同義語関係があるかを判断しなければならない。したがって実際の照合結果は、○と評価された用言を用いて質問文を拡張した評価2の結果と、○と△と評価された用言を用いた評価3の結果の間になるものと考えられる。

評価3と評価4の照合結果の差は、×と評価された用言を検索拡張に利用したかどうかの差である。提案手法では用言の同義語の候補をユーザに選抜させることで、およそ30%程度の不適切な照合結果をとりのぞくことができることを示している。

5. メーリングリストに投稿されたメールを利用した検索システム

作成したシステムは、

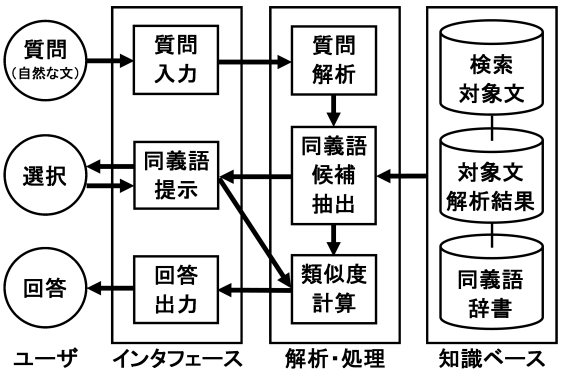

図 5　システムの概要

（1） ユーザの入力した質問文で用いられている用言の同義語候補を検索対象の文で用いられている用言から取り出す。

（2） 検索対象の文から取り出した用言の同義語候補をユーザに示し、同義語として適切なものを選ばせる。

（3） ユーザに選ばれた用言を用いて質問文を拡張し、係り受け関係の照合を行う。

図 5 に作成したシステムの概要を示す。システムを構成するモジュールの機能と内容を以下に示す。インタフェースには web ブラウザを用いた。

質問入力モジュール　自然な文で表現されているユーザの質問を受けつけ、質問解析モジュールに送る。

質問解析モジュール　ユーザの質問文を対象に形態素解析および係り受け解析を行い、解析結果を類似度計算モジュールに送る。形態素解析には JUMAN [6]、係り受け解析には KNP [7] を用いた。

同義語候補抽出モジュール　3. 章で述べた手法で、ユーザの質問文で用いられる用言と検索対象の文で用いられる用言の類似度 (同義語らしさ) を計算し、上位 5 つまでの情報を同義語提示モジュールに送る。

同義語提示モジュール　同義語候補抽出モジュールが取り出した同義語候補をユーザに示し、同義語として適切なものを選ばせる。ユーザが指定した用言は類似度計算モジュールに送る。

類似度計算モジュール　ユーザの拡張された質問文と問合わせメールから取り出した文の類似度を、文の構文的な構造と単語の重要度にもとづいて計算する [5]。結果は回答出力モジュールに送られる。

検索対象文　Vine Users ML に投稿された問合わせメールとその回答メールが格納されている。

検索対象文の解析結果　問合わせメールから取り出した問合わせの中心になる文とその前後の文の形態素解析結果および係り受け解析結果が格納されている。この解析結果は類似度計算を行うときに参照される。

同義語辞書　類似度計算で用いる同義語の辞書。名詞を中心に 519 語が登録されている。

回答出力モジュール　類似度計算モジュールの計算結果にしたがって、ユーザの質問に類似すると判定した問合わせのメールとその回答のメールを出力する。

参 考 文 献

[1] 斎藤, 大嶽, 木村: “日本語文書検索における、シソーラスによるクエリー拡張効果の分析” 電子情報通信学会技術研究報告, NLC2001-7, pp. 41-48, (2001).

[2] 太田, 奥村: “EDR 電子化辞書を用いたクエリー拡張による検索支援” 言語処理学会第 3 回年次大会, A4-6, pp. 373-376, (1997).

[3] 上野, 森, 木戸, 中川: “係り受けの 2 部グラフと共起関係を利用した同義表現抽出”, 情報処理学会研究報告, 2004-NL-159, pp.169-176, (2004).

[4] 鍜治, 河原, 黒橋, 佐藤: “格フレームの対応付けに基づく用言の言い換え” 自然言語処理, Vol.10, No.4, pp65-81, (2003).

[5] 渡辺, 横溝, 西村, 岡田: “メーリングリストを利用した質問応答システムのための知識獲得.”, 自然言語処理, vol.12, No.6, pp.25-44, (2005).

[6] 黒橋, 長尾: “日本語形態素解析システム JUMAN version 5.1 使用説明書.”, 京都大学, (2005).

[7] 黒橋: “日本語構文解析システム KNP version 2.0 使用説明書.”, 京都大学, (2005).