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

2006－NL－176 (10)
2006／11／22

# 話題語を手がかりとしたブログからのイベントマイニングの検討

数原 良彦[†] 戸田 浩之[††] 櫻井 彰人[†,†††]
[†]慶應義塾大学大学院 理工学研究科 開放環境科学専攻
〒223-8522 神奈川県横浜市港北区日吉 3-14-1
[††]日本電信電話株式会社 NTT サイバーソリューション研究所
〒239-0847 神奈川県横須賀市光の丘 1-1
[†††]CREST, 科学技術振興機構
E-mail: [†]{suhara, sakurai}@ae.keio.ac.jp, [††]toda.hiroyuki@lab.ntt.co.jp

あらまし 本研究では，関連する 2 つの話題語を手がかりとして，動作主，動作対象，動作を記述する動作関係を抽出する手法を提案する．ここでいう話題語とは，最新のブログ記事中で特徴的に出現する固有名詞のことで，従来手法によりブログ記事から自動的に抽出する．我々は，関連する話題語間の関係を記述することで，より的確に話題の内容について記述できると考える．提案手法では，2 つの話題語をキーワードとして行った AND 検索により取得したブログ記事から，キーワードに付属する格助詞とそれに伴う述語を用いて関係を記述する．しかし，実際にはしばしば省略が行われるため，2 つのキーワードが共起する文を見つけることが難しい．そこで，与えられた 2 つのキーワードについて，ひとつのキーワードが現れる文の述語と，もうひとつのキーワードが現れる文の述語が同一の場合，これは一つの動作関係を表しているとして，動作関係の抽出を実現する．

キーワード 情報抽出，関係抽出，Web マイニング，テキストマイニング

# Event mining from Blogsphere using topic words

Yoshihiko SUHARA[†] Hiroyuki TODA[††] Akito SAKURAI[†,†††]
[†]Graduate School of Science and Technology, Keio University
Hiyoshi 3-14-1, Kouhoku-ku, Yokohama, Kanagawa, 223-8522 Japan
[††]NTT Cyber Solutions Laboratories, NTT Corporation
1-1, Hikarino-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan
[†††]CREST, Japan Science and Technology Agency
E-mail: [†]{suhara, sakurai}@ae.keio.ac.jp, [††]toda.hiroyuki@lab.ntt.co.jp

**Abstract** In this paper, we propose a method to extract "action relations" between related topic words from Japanse weblog (blog). The action relation is a tuple of an agent, a target and a predicate. Our method obtains blog articles that contain two keywords by AND search and outputs action relations constructed from the predicate and the two keywords with following case particles. A noun which is followed by a case particle is called a case element. However, since words are often omitted after once they appear, the sentences with the targeted keywords are scarce. Our method solves this problem by combining two sets of sentences, each of which contains a keyword and a predicate common with the other.

**Keyword** information extraction, relation extraction, web mining, text mining

## 1．はじめに

近年のブログの普及に伴い，ブログ検索サービスの必要性が高まっている．最新の情報がすぐに更新される，ブロガーの興味を反映するという理由から，ブログ記事には有益な情報が多くあると考えられ，その抽出への取り組みがなされている．例えば，ブログ記事中の評判情報に注目したサービスとして，BuzzTunes[1]，BlogWatcher[2]などがある．また，ブログから抽出したキーワードを話題語として表示する取り組みもある．しかし，ひとつのキーワードだけでは，話題になっているイベント自体についての十分な情報が得られない．イベントとは，文

1 http://www.bztns.jp/
2 http://blogwatcher.pi.titech.ac.jp/

書中で表現されている実世界の出来事や動作のことを示す．BLOGRANGER[3]，kizasi.jp[4]などでは，共起頻度の高いフレーズや固有名詞を関連のあるものとしてグループ化し，提示している．これをもとにユーザがイベントを推定することができる．それでも，ユーザが事前知識を持たない限り，提示されたキーワード群からだけではイベントを推測することは難しい．例えば，「イスラエル」「レバノン」が関連する話題語として提示されても，事前知識のないユーザは，現実で起こっているイベントについても内容を知ることができない．

そこで，我々は関連するキーワード間の関係を抽出し，提示することで，より的確にイベントを表現する事ができるのではないかと考えた．

ここで言うキーワード間の関係とは，動作，所属，役割，位置，社会的関係などである．本研究では，これらの関係の中でも，イベントを最もよく表現すると考えられる動作関係に注目し，関連する話題語間の動作関係をブログ記事から抽出することとした（図 1）．このように，イベントの内容を簡潔な表現として抽出することをイベントマイニングと呼ぶ．

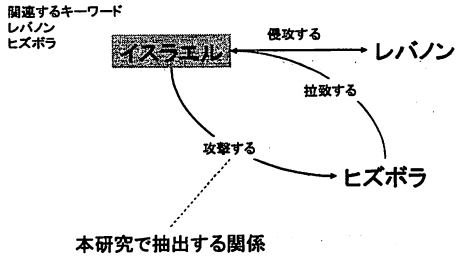


図 1: 本研究において抽出する動作関係

本研究では，動作関係を抽出するために，述語と，その述語に係る格要素の組み合わせに注目する．ここで，格助詞を伴う名詞のことを格要素と呼ぶ．しかし，前の文に現われた語は，しばしば指示詞に置換されたり，省略されたりする．省略された格要素のことをゼロ代名詞と呼ぶ．ひとつの文から格要素の組み合わせによって動作関係を抽出するためには，ゼロ代名詞を照応解消する必要がある．ゼロ代名詞の照応解消処理には，ゼロ代名詞出現箇所の「検出」と指示対象の「特定」が必要である[1]．既存の照応解消手法の多くは，ゼロ代名詞があらかじめ正しく検出されているという前提で，指示対称の特定だけに注目している．ゼロ代名詞出現箇所の検出も考慮した研究では高い精度を得てはいない．[1]．

本研究においては，ブログ記事という大量のデータを利用することができるため，上記の問題を解決するために，(i)データの中から解析可能な文を見つける，(ii)不完全な文を大量に解析して情報を補完する，という 2 つのアプローチが考えられる．2 つのキーワードが格要素として共起する頻度が少ないという予備実験の結果を踏まえ，本研究では，(ii)の不完全な文を大量に解析することで不明な格要素を補完し，動作関係を抽出する手法を提案する．

## 2. 関連研究

キーワード間の関係抽出は，これまで Message Understanding Conference (MUC)を中心に行われてきた情報抽出と見なすこともできる．MUC の課題では，文書から抽出すべき情報が，あるシナリオに沿ったテンプレートと呼ばれる一種のフレーム構造で定義されており，そのテンプレートに関する情報抽出を行っている．シナリオを限定することで，それに特化した辞書やパターンを作成することが可能となり，高い精度が得られる．

格助詞を用いた正規表現パターンを対象となる文に適用することで，特定の情報を抽出する取り組みがある．倉島ら[2]は，ブログ記事から地域に関する個人の体験情報を抽出している．また，藤井ら[3]は，百科事典の用語の説明文を Web ページから抽出している．桜井ら[4]は，ユーザに入力された用語に対する説明文を動的に Web から収集，編集してユーザに提示するシステムを提案している．

Hasegawa ら[5]は，文中に 2 つの固有表現が現れた場合，その出現の間にある文字列がこれらの固有表現の関係を表す可能性が高いと考えた．この文字列によってクラスタリングを行い，文字列中の頻出単語をラベルとして，人と組織，企業と企業の president, M&A, rival といった静的な関係を新聞記事から抽出している．また，森ら[6]は，関連する固有表現が与えられたという条件の下で，それらの固有表現を AND 検索のキーワードとして Web ページを取得し，Hasegawa らと同様に固有表現の間に現れた文字列を利用して，所属，役割，位置

3 http://ranger.labs.goo.ne.jp/

4 http://kizasi.jp/

などの関係を抽出している．本研究は，抽出する関係を特定しないという点でこれらの研究とは異なる．

関係抽出の問題では，教師付き学習を用いた手法もいくつか提案されているが[7]，その場合には，キーワードの組み合わせを限定する必要があるという短所がある．また，本研究では，抽出する関係を特に限定しないため，教師データを作成するのが困難となり，本課題に教師付き学習を適用するのは難しいと考える．

## 3. 格助詞を用いた動作関係の抽出

### 3. 1 格助詞の選択

本研究では，ガ格，ニ格，ヲ格に限定して動作関係の抽出を行う．また，句点を区切りとして文書を文に分割する．日本語では，格要素と述語が出現することで，文内の格関係を形成する．ガ格は主に運動の主体，ニ格は間接的な対象など，ヲ格は動作の直接的な対象などの役割を表している[8]．この格の役割は，格要素の名詞や，その他の格要素，述語との関係で意味が変わる．本研究では，単純にガ格を「動作主」，ニ格，ヲ格を「動作対象」，述語を関係の記述と見なして関係抽出を行う．例えば，

(1) イスラエル が ヒズボラ を 攻撃する
(2) イスラエル が レバノン に 侵攻する

という文の場合，(1)では動作主が「イスラエル」，動作対象が「ヒズボラ」で，その動作関係は「攻撃する」となる．(2)の場合は，動作主が「イスラエル」，動作対象が「レバノン」，動作関係は「侵攻する」と解釈する．

係助詞「は」は，文の部分を強調するために，他の格助詞の代わりに用いられる[8]．本研究では，「は」はガ格の格助詞とみなす．

### 3. 2 受動態の処理

受動文では格の意味役割が変化する．一般的に受動文の場合，ガ格が動作対象を表し，ニ格が動作主を表現する．

(3) ヒズボラ が イスラエル に 攻撃される
(4) ヒズボラ に イスラエル が 攻撃する

本研究では，動詞の未然形に接尾辞「れる」が連接するものが述語として抽出された場合，その文を受動態と判別し，ガ格とニ格を交換し，述語を原形に変換する．

### 3. 3 中央埋込文と重文の処理

中央埋込文 (例1)や重文(例 2)には，ひとつの文に複数の述語が出現する．提案手法では，以下に示すように，中央埋込文では埋め込まれた文の述語を，重文では前の文の述語を抽出する．

(例1) 太郎は [花子が 次郎を 好き] だと 思っている
$S_1$ $S_2$ $V_2$ $V_1$

(例2) [おじいさんは 山へ行き] おばあさんは 川へ行った
$S_1$ $V_1$ $S_1$ $V_2$

### 3. 4 格要素の文内共起

3.1 節で述べたような関係の抽出を行うためには，ひとつの文の中に，目的の 2 つのキーワードが格要素として出現する必要がある．

そこで，関連する 2 つのキーワードが，(A)格助詞を伴って同一文内に共起する文の頻度，(B)少なくとも一方のキーワードが現れる平均文数を計算する予備実験を行った．13 組のキーワードでそれぞれ AND 検索を行い，50 語(500 語)のスニペット 500 件ずつ取得し，500 件に含まれる平均文数を計算した．総文数平均は 500 件に含まれる総文数の平均を表す．実験の結果を表 1 に示す．

表 1: キーワードペアの文内共起の平均頻度

| | 50 語 | 500 語 |
|---|---|---|
| 総文数平均 | 1116.9 文 | 5008.2 文 |
| (A) 共起 | 6.8 文 | 17.6 文 |
| (B) どちらか | 105.6 文 | 161.7 文 |
| (A) / (B) | 6.4% | 10.9% |

このうち，スニペット長 500 語では 4 組のキーワード，50 語では，6 組のキーワードにおいて，格要素が文内共起する文が 1 つも取得できなかった．

## 4. 提案手法

3.4 節で行った予備実験の結果を踏まえ，本研究

では，キーワードひとつずつの解析によって動作関係の抽出を行う手法を提案する．あるキーワードが格助詞を伴って現れた文の，キーワード・格助詞・述語をひとつのパターンとして抽出したものを述語パターンと呼ぶ．提案手法では，同一の述語を持つ述語パターンを組み合わせることによって格要素の補完を行う．補完されたものを動作パターンと呼ぶ．

提案手法は，以下の 4 ステップから成る(図 2)．

1. 関連する 2 つのキーワードの取得
2. ブログ記事の取得
3. 述語パターンの抽出
4. 動作パターンの抽出

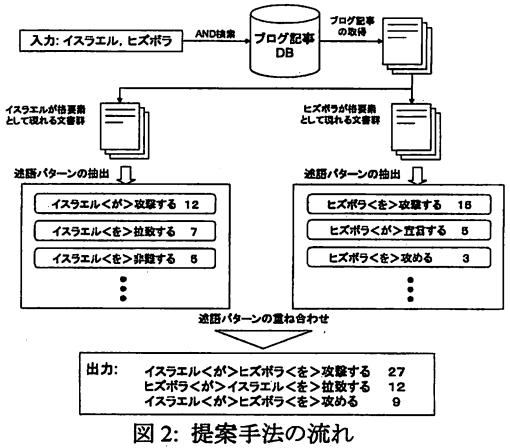


図 2: 提案手法の流れ

### 4. 1 関連する 2 つのキーワードの取得

関連する 2 つのキーワードは，BLOGRANGER[5]の”最新の注目キーワード”ページから，関連付けられた 2 つの話題語を取得する．BLOGRANGER の話題語抽出と関連付けのアルゴリズムは以下の通りである．

(1) 最新記事のうち、ブログの注目度を表す Eigenrumor スコア[9]が閾値以上の 10,000 記事を選択する．
(2) 選択された記事から固有表現を抽出する[10]．
(3) 最新記事の中で特徴的に存在する固有表現を話題語として抽出する[11]．
(4) 抽出した固有表現のうち，出現する文書の重なりが閾値以上の組み合わせを関連性がある固有表現ペアとして特定する．

### 4. 2 ブログ記事の取得

関係抽出を行うためのブログ記事は，2 つのキーワードの AND 検索で行う．まず，AND 検索を用いて 2 つのキーワードを含む文書群を取得する．

### 4. 3 述語パターンの抽出

取得したブログ記事は，まず句点を区切り文字とした文に分割する．そして，各文に対して MeCab[6]を用いて形態素解析を行う．

格助詞は，「が・は」をガ格，「に」をニ格，「を」をヲ格として，その直前の名詞とあわせて格要素とする．また，格要素が現れてから最初に現れる動詞を述語として抽出する．動詞「する」に関しては，直前にサ変接続名詞が連接している場合，その名詞を結合し，サ変動詞とする．

格助詞の直前の名詞として，最長連接名詞を用いる．この最長連接名詞にキーワードが含まれていれば，述語パターンとして抽出する．最長連接名詞で判定を行っているため，ある程度の表記ゆれに対応することができる．また，動作を表さない「する，ある，なる，いる，やる」は不要語として除去する．

### 4. 4 動作パターンの抽出

2 つのキーワードに対応する述語パターンについて，同一の述語を持つものを選択し，組み合わせて動作パターンとする．この際，同じ格助詞同士，ニ格とヲ格の組み合わせは行わない(図 3)．動作パターンはスコア順に出力される．このスコアは，(i)述語パターンの頻度の和，(ii)頻度の和を TF-IDF 重みづけをしたもの，を用いる．TF-IDF には以下の値を用いる．

$$tfidf = \log(tf)\cdot\left\{\log\left(\frac{N}{df}\right)+1\right\}$$

## 5. 評価

抽出された動作関係が適切であるか，ベースライン手法と比較実験を行った．また，3.2 節で述べた受動文の取り扱いに関する提案手法，すなわちガ格とニ格を入れ替える手法の評価も一部行った．

[5] http://ranger.labs.goo.ne.jp/br1/

[6] http://mecab.sourceforge.jp/

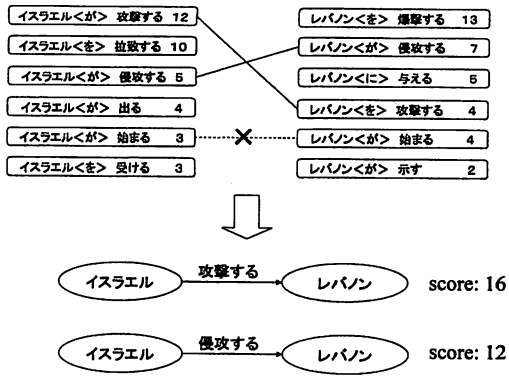


図 3: 動作パターンの抽出

**イスラエル—レバノン**

| | |
|---|---|
| イスラエル<が>レバノン<を>攻撃する | 16 |
| イスラエル<が>レバノン<に>侵攻する | 12 |
| イスラエル<が>レバノン<を>撤退する | 8 |

**イスラエル—ヒズボラ**

| | |
|---|---|
| イスラエル<を>ヒズボラ<が>拉致する | 12 |
| イスラエル<が>ヒズボラ<を>攻撃する | 8 |
| イスラエル<が>ヒズボラ<を>攻める | 4 |

**スーパーモーニング—亀田史郎**

| | |
|---|---|
| スーパーモーニング<を>亀田史郎<が>出演する | 4 |
| スーパーモーニング<に>亀田史郎<が>出演する | 4 |
| スーパーモーニング<に>亀田史郎<が>出る | 2 |

図 4: 出力結果の例

### 5.1 実験データ

ブログ記事の検索には，BLOGRANGER API[7]を利用した．この API を利用することで，日本語で記述されたブログ記事のうち，直近の 1 ヶ月間に更新されたブログ記事約 1,000 万件を対象に検索を行うことができる．検索結果のランキングには，Eigenrumor スコアを利用した．10 組のキーワードについて AND 検索を行い，$n$ 語長のスニペットを 500 件取得した($n$=50, 500)．ここでいうスニペットとは，該当キーワードを出来るだけ多く含む連続する $n$ 語の文字列のことを表す．API の仕様上，2 つのキーワードを含むスニペットを取得するように指定することができないため，ひとつのキーワードを含む条件での検索結果を 500 件，もう一方のキーワードを条件に含めた検索結果を 500 件取得した．すなわち，1 つのキーワードの組に対して 2 回検索を行い，1,000 件のスニペットを取得した．

[7] http://ranger.labs.goo.ne.jp/hacks

### 5.2 比較手法

提案手法と比較を行うため，取得した検索結果の中から，2 つのキーワードが格要素として共起する文を選択し，述語を動作関係として抽出する手法をベースライン手法とした．

### 5.3 評価指標

抽出された動作関係が適切なものかを判定するために，以下の指標を用いた．受動態を変換したものは，最上位の正解率のみ比較を行った．

(a) 最上位の正解率

スコアが最も高い動作関係の正解率．

(b) 上位 3 つの正解率

スコア上位 3 つの動作関係の正解率．

(c) 平均逆順位

スコアが最高順位である正解の動作関係の平均逆順位．

### 5.4 実験結果

最上位の正解率を図5，上位3つの正解率を図6，正解の平均逆順位を図 7 に示す．

3 つの評価指標において，提案手法がベースライン手法を上回った．最上位の正解率において，2 つの手法ともにスニペット長が 50 語の方が 500 語よりも精度が高いことがわかる．他の 2 つの評価指標についても，同様のことがいえる．また，提案手法では，スコアを TF-IDF 重みづけしたもの(tfidf)が，重みづけなし(normal)とほぼ同じ値を示しているが，ベースライン手法では，TF-IDF 重みづけを行ったものが，精度が低い．

また，受動文の変換を行うことで，ベースライン手法では最上位の正解率が低くなっているものの，提案手法ではそれほど変化していないことがわかる．

### 5.5 検討

結果より，スニペット長が 500 語より，50 語の方が 2 手法において精度が高いことがわかる．これは，文書中の広い範囲から文を取得することで，異なる話題で当該キーワードが出現する文を含んでしまう可能性が高いためだと考えられる．例えば，イスラエルによるレバノン侵攻の話題の前後には，侵攻の背景などが書かれていることがある．このような文に含まれる述語パターンがノイズとなり，精度を低下させているのだと考えられる．

ベースライン手法で受動文変換を行ったもので

は精度が低い．これは受動文の誤った判別・変換がされてしまったことによるものだが，提案手法がベースライン手法より，ロバストな性質を持っていることが伺える．

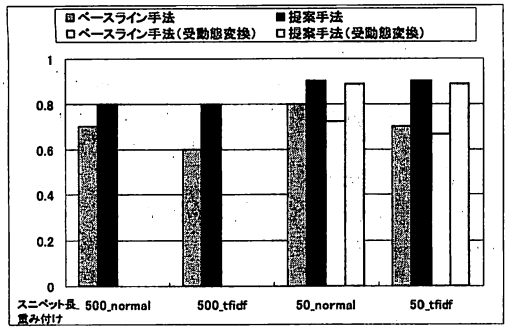


図 5: 最上位の正解率

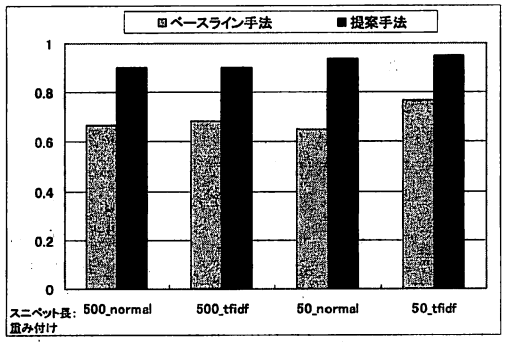


図 6: 上位 3 つの正解率

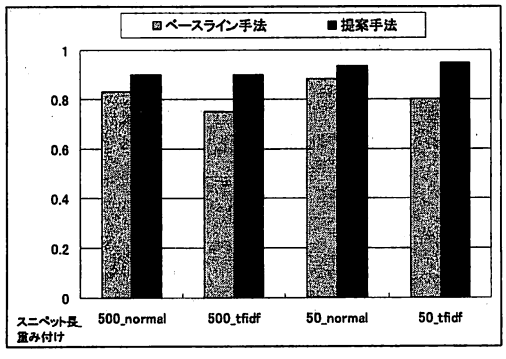


図 7: 正解の平均逆順位

## 6. 今後の課題

本研究での評価実験は，量が不足しているため，実験データの量を増やす必要がある．また，受動文を変換する手法の評価については，さらに実験を行った後で，検討を行う必要がある．

受動文の判別規則を改善する必要がある．これについては，機械学習を用いて，受動文・使役文を非常に高い精度で能動文へ書き換える手法が研究されている[12]．本研究では，訓練データを用意することが困難ではあるが，他のコーパスで学習した分類器を利用することも考えられる．

本研究では，ブログ記事から抽出された話題語を用いたため，関係抽出を行う対象もブログ記事に限定していた．同じ手法を新聞記事に適用するつもりである．

## 7. まとめ

評価実験の結果，提案手法によって，2 つのキーワードが格助詞を伴って文内共起する文を解析するベースライン手法を上回る精度で動作関係を抽出することが確認できた．今後は追加実験を行うことで，提案手法の改善を行う予定である．

動作関係の抽出は，関係抽出の研究でも殆ど研究されていない分野である．本手法と他の関係抽出手法と組み合わせることで，更に幅広い情報抽出が可能になると考えている．

## 参考文献

[1]関和広，藤井敦，石川徹也．ゼロ代名詞の検出と補完を統合した確率的照応解消モデル．言語処理学会第 8 回年次大会発表論文集，pp.591-594. (2002).

[2] 倉島健，手塚太郎，田中克己．Blog からの街の話題抽出手法の提案．電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005) (2005).

[3] 藤井 敦，渡邉 まり子，石川 徹也．事典的 Web 検索サイトにおける複数文書要約の応用．言語処理学会第 10 回年次大会発表論文集，pp.261-264. (2004).

[4] 桜井裕，佐藤理史．ワールドワイドウェブを利用した用語説明の自動生成，情報処理学会論文誌,Vol.43,No.5,pp.1470-1480. (2002).

[5] Hasegawa, T., Sekine, S. and Grishman, R. Discoverying relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting of the ACL*. (2004).

[6] 森純一郎，辻下 卓見，松尾 豊，石塚 満．

Web からのエンティティ間の関係情報の抽出. 第 20 回人工知能学会全国大会 (JSAI2006). (2006).

[7] Zelenko, D., Aone, C. and Richardella, A. Kernel methods for relation extraction. *Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for ACL*, pp. 71-78. (2002).

[8] 高橋太郎. 日本語の文法. ひつじ書房. (2005)

[9] Fujimura, K., Inoue, T. and Sugisaki, M. The EigenRumor Alogorithm for Ranking Blogs. In *Proceedings of the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (2005).

[10] Isozaki, H. and Kazawa, H. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics*, pp.1-7. (2002).

[11] Toda, H. and Kataoka, R. A search result clustering method using informatively named entities. In Proceedings of the 7th Annual ACM international Workshop on Web information and Data Management, pp.81-86. (2005).

[12] 村田真樹, 井佐原均. 受け身／使役文の能動文への変換における機械学習を用いた格助詞の変換.情報処理学会自然言語処理研究会, 2002-NL-149. (2002).

[13] 徳永健伸. 情報検索と言語処理, 言語と計算 5. 東京大学出版会. (1999).