

## 統計的に獲得された用語への属性ラベル付与

池野 篤司<sup>†</sup> 濱口 佳孝<sup>†</sup> 山本 英子<sup>‡</sup> 井佐原 均<sup>‡</sup>

<sup>†</sup> 沖電気工業株式会社 〒541-0053 大阪市中央区本町 2-5-7 丸紅大阪本社ビル 4F

<sup>‡</sup> 独立行政法人 情報通信研究機構 〒619-0289 京都府相楽郡精華町光台 3-5

E-mail: <sup>†</sup> {ikeno546, hamaguti662}@oki.com, <sup>‡</sup> {eiko, isahara}@nict.go.jp

あらまし 情報抽出をベースとしたアプリケーションに利用される(固有表現) タグ用の辞書作成のために、Web ページなどの大規模文書集合から統計的に獲得された用語に属性ラベルを付与する試みについて報告する。与えられた用語リストを、まず基本的な属性拡張ルールを適用することにより、(1)初期属性採用語、(2)末尾属性採用語、(3)専門用語属性採用語、(4)未付与の4つのグループに分割した。その結果、約半数が属性に関する決定的な情報を持たないために未付与のグループに残った。そこで、すでに属性が付与された用語のうち、専門用語と判定されたものの構成要素を「属性影響語」と見なし、専門用語属性を仮設定して、未付与グループの用語リストに再度属性拡張ルールを適用したところ、未付与グループのさらに半数に対して属性ラベルを付与できる可能性があることがわかった。

キーワード 情報抽出、固有表現、属性ラベル、属性影響語、辞書

## Label Attachment to Statistically Acquired Terms

Atsushi IKENO<sup>†</sup> Yoshitaka HAMAGUCHI<sup>†</sup> Eiko YAMAMOTO<sup>‡</sup> Hitoshi ISAHARA<sup>‡</sup>

<sup>†</sup> Oki Electronic Industry Co., Ltd.

4F Marubeni Building, 2-5-7 Honmachi, Chuo-ku, Osaka-shi, Osaka, 541-0053 Japan

<sup>‡</sup> Independent Administrative Institution, National Institute of Information and Communications Technology

3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan

E-mail: <sup>†</sup> {ikeno546, hamaguti662}@oki.com, <sup>‡</sup> {eiko, isahara}@nict.go.jp

**Abstract** This report describes an experimental result of attaching labels to statistically acquired terms from the Web pages, for the purpose of using the terms as dictionary entries to improve the accuracy of information extraction in a practical application. A terms list is divided into 4 groups, consisting (1) Initial Label Adopted, (2) Last Label Expanded, (3) Technical Term Label Expanded, (4) Not Attached, by applying basic label expansion rules. Since the result of the experiment shows that half the terms are assigned to “Not Attached” group, we have conducted additional experiment on the term list of the group, which is to detect the technical term label affectants, replace their labels temporarily, and re-apply the expansion rules. The result shows that almost half the terms in “Not Attached” group are able to belong to the other groups.

**Keyword** Information Extraction, Named Entity, Label Attachment, Label Affectant, Dictionary

### 1. はじめに

我々は情報検索と情報抽出を応用したアプリケーションとして「産学連携支援ツール Bluesilk<sup>®</sup>」[1,2]の開発に取り組んでいる。これは論文・特許・Web などから、所望の情報をピンポイントで得るためのシステムである。主な改良項目として、情報検索に関する精度向上にも取り組んでいる[3]が、情報抽出に関連する部分の精度向上も重要な課題である。

情報抽出においては、まず固有表現(NE)タグ付けが行われるのが一般的である。NE タグ付け器(以下、NE タガーと呼ぶ)には大きく分けて、文字ベースのもの、形態素解析ベースのものがある。我々は固有表現以外のラベルをつけることができるように NE タガーの拡張を行っており、その際の有利さから形態素解析をベースとする手法を用いている。

このとき問題となるのが、形態素解析ベースの NE タガーは形態素の切り方を間違った場合の精度の低下が大きいことである。特に Web のように日々新語・造語が現れる文書を処理対象にする場合には、それらの

<sup>1</sup> Bluesilk<sup>®</sup>は(株)三菱総合研究所の登録商標である。

未知語による解析誤りは NE タガーにとっては致命傷となりうる。

この問題を解決する最も簡単な方法は、形態素辞書を追加することである。この方法であれば NE タガーの動作速度にも影響を与えることがないというメリットもある。

そこで我々は、Web ページを収集した大規模文書集合から、未知語や専門用語を統計的に獲得し、さらにそれらに対して NE タガーが利用できる属性（固有表現）ラベルを付与することで辞書化することを考えた。前半の用語の統計的獲得に関しては文献[4]に詳細が記されているので参照されたい。本稿では、後半部分の、獲得された用語に対して、NE タガーで利用するための属性ラベルを付与する手法および実験について報告する。

ラベル付与の際に注意しなければいけないのは、継続運用する応用アプリケーションに対して最新の辞書データを追加していくことを想定しているの、精度を保ちながらも処理全体をできる限り単純・高速なものにする必要がある点である。そのため、本研究で用いる手法では、用語リストだけを受け取り、原文に関する情報は利用せずに処理することを試みる。

なお、本研究で用いる NE タガーは固有表現だけでなく技術名などにもタグ付けするように拡張しているが、簡単のため、本稿では拡張した NE タガーを単に NE タガーと記述する。人名や技術名などのいわゆる固有表現ラベルに相当するものについては属性ラベルという表記で統一するものとする。

2 章でまず応用アプリケーションとして想定している Bluesilk<sup>®</sup>について説明した後、3 章で属性ラベル付与手順を述べる。4 章で実際の実験結果とそれを踏まえた検討を記し、5 章で実験全体を考察する。

## 2. Bluesilk<sup>®</sup>について

### 2.1. 概要

Bluesilk<sup>®</sup>は、(技術的な内容の)テキストを入力すると、Web ページや論文・特許などの文書集合から要求内容に関連する文書を検索し、さらに指定された属性を持つ語だけを抽出してリストアップすることができるシステムである。

システムの目的は、産学が効率的に連携するための支援をすることである。そのために、連携相手を効率的に探せること、提供すべき情報の入力を容易にすること、の 2 点を目指して、沖電気工業株式会社と株式会社三菱総合研究所により共同開発中である。

プロトタイプでは、キーパーソン抽出(「人名」)、企業・組織抽出(「組織名」)、技術抽出(「技術名(相当語)」)などの指定ができるようになっている。

利用場面としては、企業が必要とする技術を保持する大学の研究者を探したり、大学の研究者が自分の研究を製品化することができる企業を探したり、といったものを想定している。

### 2.2. 特徴

Bluesilk<sup>®</sup>の特徴は、情報を検索する場面と情報を提供する場面の双方にある。

#### 2.2.1. 情報の検索

一般的な情報検索(文書検索)とは以下の点で異なっている。

##### (1) 情報検索のピンポイント化

検索エンジンでの「検索」とは、入力した単語を含む文書の所在を(せいぜい関連する要約文を付与して)示すだけである。それに対して Bluesilk<sup>®</sup>では、文書検索に加えて情報抽出の手法を導入し、検索結果の文書の中から、人名・組織名などの必要な情報のみをピンポイントに抽出できる。

##### (2) 抽出語の統計処理

ピンポイントで抽出された語は、出現文書・出現位置・出現回数などの統計量を利用して集計・順序付けして提示する。これにより、ユーザは抽出された人名や組織名などが、どれくらい入力した内容に関連しているかの指標を得ることができる。

#### 2.2.2. 情報の提供

大学で産学連携が言われるようになった初期から、研究内容を研究者自身に専用フォーマットに入力させたものをデータベース化する、という試みが行われている。しかし、研究内容を改めて専用のフォーマットに入力するという作業は研究者にとって負担になることも多く、協力が得られなかったり、一度入力した後はほとんど更新されなかったりと、情報提供に対する敷居が高いのが現状である。以下にこの問題に関係する Bluesilk<sup>®</sup>の特徴を記す。

##### (1) 非定型文書からのデータベース作成

Bluesilk<sup>®</sup>に取り入れられる文書は定型フォームに依存しない。つまり、データベースへの登録を前提とした書式にあわせて情報を入力する必要がないのである。研究者が普段の業務の中で作成した文書を提供するだけで、専用フォーマットに入力された文書から作成されたデータベースとほぼ等しいものが生成できる。これは、先に検索に関して挙げた情報抽出の技術により実現している。

登録するのは論文・報告書・講義メモなど、どのような種類の文書でもよいが、我々が特に着目したのは Web ページである。昨今は大学・企業とも Web ページにより情報を発信することが多くなってきており、Web ページから情報を抽出してデータベースに登録することができれば利用価値が高まると

考えて Bluesilk® に機能として取り込んだ。

## (2) 即時性

Bluesilk® では、既存の文書データの他に、インターネットやイントラネット上の最新情報を自動収集・解析してコンテンツとして取り入れる機能を有している。このため、常に最新情報を対象に情報を抽出することが可能となる。研究者が手元に保持する文書を更新すれば、データベースの内容も自動的に更新されることになる。

従来のデータベース方式では、データ収集から書式の変換、抄録やキーワードの付与作業等に数ヶ月以上の時間が経過してしまうこともあるため、最新情報を取得するには制限が生じていた。

## 2.3. 構成

システムは大きく、データを収集・整理する部分と、ユーザからの情報探索要求に対応する部分との2つに分かれる。さらにそれぞれが、データ収集部と情報抽出部、文書検索部と統計処理部に分かれる。

### (1) データ収集部

インターネット・イントラネット上の Web ページを一定期間ごとに収集し、検索対象文書としてデータベースに登録する。このとき文書検索や情報抽出のために HTML 形式の Web ページを加工する。論文・特許などの静的な文書はここで受け付けて、そのまま検索対象文書としてデータベースに登録する。

### (2) 情報抽出部

収集された文書に対して、固有表現（人名、組織名、など）を中心とした各種の語を抽出し、抽出情報データベースに登録する。定型文書に対しては文書フォーマットの情報を利用して抽出の精度を高める。一方で、Web ページのような非定型な文書に対しては、その構造情報を利用して精度を高める。

### (3) 文書検索部

検索対象文書データベース中の文書のインデックスを随時作成しておく一方で、ユーザからの情報探索要求の入力に対しては、インデックスを用いて関連文書を検索する。

検索要求は形態素解析されるため、キーワードだけでなく自然文で入力することが可能である。基本的には検索要求中の形態素が多く含まれる文書が関連文書となるが、検索要求中の形態素そのものが含まれる文書を検索する全文検索と、検索要求中の形態素の関連語（全文検索の結果得られた文書中の主要語）を用いて文書を検索する概念検索を用意している。

形態素解析には chasen[5]、検索のコアエンジンには GETA[6]を用いている。産学連携を主目的に掲げ

ていることもあり、技術用語に関しては独自に語彙を増強している。

### (4) 統計処理部

入力時に指定された属性（人名、技術名など）の語について、統計情報をもとにスコアリングした結果により順位付けをして出力する。利用する情報は、出現頻度・検索要求中の形態素あるいは関連語との距離などである。それぞれの情報に一定の重みパラメータを乗じて評価関数を構成している。パラメータの値は実験により最適になるように調整している。

## 3. 属性ラベル付与の概要

獲得された用語を情報抽出のための辞書データとして用いるために、本研究では、与えられた用語リストに対して属性ラベルを付与することを試みる。

属性ラベル付与のための基本情報としては、技術名などの専門用語も抽出可能にした NE タガーを用いて付与した初期属性ラベルを利用する。

NE タガーの既存データ・既存ルールにより用語全体に属性ラベルを付与されれば、それらの語はその属性を使ってそのまま辞書データとすることが可能である。

一方で、NE タガーでは用語全体の属性ラベルが付与できない場合もある。このとき、できる限り多くの語について適当な属性ラベルを推定・付与して辞書データとする方法を検討し、実験を行う。

実験は、用語のいずれかの構成要素の属性ラベルを全体に反映させる単要素属性拡張をまず行う。次に、それでも属性ラベルが付与できなかった用語を対象として、再度拡張を試みる。このとき、単要素で拡張できた用語の構成要素から選択した属性影響語を用いる。以下に実験の詳細を記す。

### 3.1. 単要素属性拡張

与えられた用語リストを、属性拡張ルールを適用することにより以下の4つのリストに分割する。ルールが適用されなかった用語についてのみ次のルールを適用するため、各グループは排他的となる。

#### (1) 初期属性採用用語リスト

用語全体に (NE タガーによる初期)属性ラベルが付与されている用語をリストアップする。

#### (2) 末尾属性採用用語リスト

用語を構成する末尾の要素の属性ラベルを、用語全体に割り当てることができるものをリストアップする。

一般に日本語の複合語においては、末尾の要素が複合語全体の品詞・意味を規定すると考えられるからである。

末尾以外の要素には特定の属性ラベルが付与されていない場合には無条件で末尾要素の属性ラベルを採用する。また、末尾要素属性と同じか同類の属性ラベルを持つ要素のみが存在する場合も、末尾要素の属性ラベルを採用する。

末尾要素の属性ラベルと、他の要素の属性ラベルが異なる場合には、相互の相性を記したルールにより採用・不採用（未付与）を判断する。

### (3) 専門用語属性採用語リスト

特定の分野に属する文書群から統計的に獲得された用語は、その分野特有の専門用語を多く含むと考えられる。この仮定に基づいて、末尾以外に専門用語属性を持つ構成要素が存在する場合に、上記(2)と同様の条件で用語全体の属性ラベルを付与することができたものをリストアップする。

### (4) 未付与語リスト

上記のいずれのグループにも振り分けられなかった用語をリストアップする。

## 3.2. 属性影響語による拡張

3.1 節における(4)未付与語リストを対象に、さらに属性ラベル付与を試みる。

この時点でリストに残っている語には属性に関する決定的な情報を持たないため、既知の情報を用いて、未付与語に関連する情報を補間する必要がある。

そのため、以下の手順でいくつかの構成要素に対して属性ラベルを仮設定して属性拡張する。

#### (1) 「属性影響語」の抽出

特定の属性を選択し、その属性ラベルを付与された用語の集合から頻出する構成要素をリストアップして、それらを「属性に影響を与える語（属性影響語）」であると想定した。

#### (2) 属性の仮設定

属性影響語のうち、現在は属性ラベルを持っていない語に対して、(1)で選択した特定の属性を一時的に設定する。

#### (3) 拡張ルールの適用

属性影響語の属性を設定した状態で、未付与語リストを、3.1 節の(2)、(3)に分類する。

## 4. 実験

### 4.1. 実験データ

ここでは、東京大学の工学部・工学系研究科の Web サイトから収集したホームページ（テキストのみで約 200Mbytes）から、[4]で述べられている N-1 の手法によって獲得された 6205 語に対して処理を行った。

文書集合が工学を対象としたものであるため、専門用語属性は科学技術関連の用語に付与される。

### 4.2. NE タグー

固有表現およびその他の属性のタグ付けには、沖電気で研究開発を行っている NE タグーを用いた[7,8]。タグ付け結果は要素の重複を許す。一部の語については、形態素解析による品詞情報を返すことが可能である。

### 4.3. 単要素属性拡張実験

#### 4.3.1. 結果

グループごとに、付与したラベルの適合率を求めたものを表 1 に示す。正解の判定は人手で行った。

表 1 属性ラベルの適合率

リスト	語数	正解数	適合率
(1)初期属性採用	1190	1060	0.8908
(2)末尾属性採用	864	793	0.9178
(3)専門用語属性	612	553	0.9036
(4)未付与語	3539	-	-
合計	6205	-	-

末尾属性を採用して正しくラベルが付与できた例を以下に示す。

例 1).

倒立振子      ラベル：専門用語-技術

倒立(一般名詞) + 振子(専門用語-技術)

また、専門用語属性を採用して正しくラベル付けできた例を以下に示す。

例 2).

弾塑性解析      ラベル：専門用語-技術

弾塑性(専門用語-技術) + 解析(一般名詞)

一方、誤ったラベルを付与してしまった例には以下のようなものがあつた。

例 3).

系超伝導体      ラベル：専門用語-技術

系(一般名詞) + 超伝導体(専門用語-技術)

例 4).

原子炉実習      ラベル：専門用語-技術

原子炉(専門用語-技術) + 実習(一般名詞)

#### 4.3.2. 検討

付与されたラベルの適合率は 9 割前後と、辞書登録前の人手でのチェックは必要であるとしても、実用に耐えうる範囲に収まっていると考える。

一方で、半数以上が未付与語に分類されている。ある程度ドメインの特定された文書集合から獲得された用語であるので、これらすべてがいわゆる一般的な複合名詞であるとは考えにくい。属性ラベルが付与されるべき語がまだ多く残っていると推測される。

例 3 の用語は、用語獲得の段階で分割位置が正しく決定できなかったものである。用語獲得時の情報ではこの語を候補から排除しきれないので、ラベル付与時

に排除したいと考えて誤りと判定している。今回の実験で利用している情報では排除できないが、形態素解析の詳細情報（例えば、名詞と言っても接辞の可能性を持つものであったり、サ変名詞であったり、といった情報）を用いて排除できる可能性がある。

また、専門用語属性を採用する場合には、例2のように成功する場合と、例4のように失敗する場合が存在するが、決定的な条件の差異を見つけることができない場合も多い。これについては今後実験を継続して詳細を調査したい。

#### 4.4. 属性影響語による拡張実験

##### 4.4.1. 結果

ここでの属性影響語の認定は、専門用語属性を対象に行った。一時的に属性ラベルを付与する属性影響語候補を以下に示す。

(高頻度語) 解析, 構造, システム, 制御, 工学, 計算, 情報, 光, 特性, ファイル, 冷却, 熱, 実験, 要素, 変換, 表示, 生成, 研究, 空間, 環境, 入力, 状態, 材料, 文字, 磁場, 開発, エネルギー, 理論, 分析, 物質, 評価, 電子, 通信, ...  
 (頻度2) 連立, 列, 力学, 稜線, 溶液, 面積, 未知数, 法線, 補強, 便覧, 変形, 分子, 分割, 物理, 負荷, 不変, 表現, 発生, 排水, 年度, 認知, 二次元, 読み出し, 読み取り, 特徴, 討論, ...  
 (頻度1) 濾過, 撥, 攪乱, 歪み, 歪, 話, 論題, 炉心, 励振, 励起, 力, 領域, 量, 両面, 粒子, 流動, 理, 抑制, 熔融, 揺らぎ, 予測, 融合, 誘導, 有限, 輸送, 薬品, 名前, 無線, 未踏, ...

実際に属性ラベルを付与する語は、出現頻度により3通り用意して実験した。頻度1以上のもの556語に付与した場合の結果を表2に、頻度2以上のもの195語に付与した場合の結果を表3に、頻度3以上のもの88語に付与した場合の結果を表4にそれぞれ示す。

表2 属性影響語（頻度1以上、556語）  
を利用した場合のラベルの適合率

頻度	語数	正解数	適合率
(2)末尾属性採用	1476	975	0.6605
(3)専門用語属性	641	378	0.5897
(4)未付与語	1422	-	-
合計	3539	-	-

表3 属性影響語（頻度2以上、195語）  
を利用した場合のラベルの適合率

頻度	語数	正解数	適合率
(2)末尾属性採用	1108	781	0.7049
(3)専門用語属性	532	354	0.6654
(4)未付与語	1899	-	-

合計	3539	-	-
----	------	---	---

表4 属性影響語（頻度3以上、88語）  
を利用した場合のラベルの適合率

頻度	語数	正解数	適合率
(2)末尾属性採用	834	575	0.6894
(3)専門用語属性	440	298	0.6773
(4)未付与語	2265	-	-
合計	3539	-	-

属性影響語による拡張により正しくラベルが付与できた例を以下に示す。

例5).

ウェーブレット変換 ラベル: 専門用語-技術  
 ウェーブレット(未知語-カタカナ)  
 + 変換(一般名詞→専門用語)

例6).

マイクロマシン ラベル: 専門用語-技術  
 マイクロ(一般名詞→専門用語)  
 + マシン(一般名詞)

一方、誤ったラベルを付与してしまった例には以下のようなものがあつた。

例7).

システム研究 ラベル: 専門用語-技術  
 システム(一般名詞→専門用語)  
 + 研究(一般名詞→専門用語)

##### 4.4.2. 検討

属性影響語には、専門用語としておかしくないものも多数あるが、一方で一般名詞としてよく使われるものも多く含まれている。そのため、属性影響語による拡張を行うと、ラベルを付与できる語数は増加するが、表2や例7からわかるように無条件で適用すると適合率が下がってしまう。実用にあたっては、適当な閾値で属性影響語を取捨するか、この時点で人手によるチェックが必要になるであろう。

さらに、この手法を適用した後の未付与語リストの内容分析をする必要がある。これについては今後の課題とする。

実験全体を通じて、用語リストとNEタガーからの情報のみを用いて、約3/4の用語に対して自動で属性を付与することができたことは、今回の手法が情報抽出の辞書作成に有効であることを示している。

## 5. 考察

### 5.1. 属性影響語について

属性影響語の概念は、今まで気づかなかった重要な(「用語の属性」に影響を与える)部分文字列の発見を意味している。

今回の実験では単純に、属性を持っていない構成要素(形態素)に属性ラベルを一時的に付与するという

手法をとったが、ここにはまだ様々な選択方法を試すことが可能である。例えば、対象となっている形態素が、属性を持っている長い形態素の一部になっている場合は、さらに属性影響語としての重要度を増すように調整することも考えられる。

また、4.4.1節に例示した語を見ると、必ずしも高頻度のものが影響語にふさわしいとは限らないように思われる。今回の実験での影響語の選定にあたっては、用語リスト中での出現頻度により、閾値を一つだけ用いる手法をとったが、例えば、頻度中位のものを選択する、あるいは原文書での *tf*, *df* などの値を利用する、などの選定方法を今後試みたい。

## 5.2. 獲得用語リストの意義

本研究の目的は、用語の境界がはっきりしている場合の NE タグ付け（属性ラベルの付与）と言い換えることもできる。一般には、NE タグ付けはテキスト全体に対して、周辺語の情報を参照して行われる。テキスト全体を対象としたときには、情報は豊富にある代わりに各々の語の境界が不明確であるという問題もある。タグ付け対象を、NE タガーとは別の方法で境界を明らかにした文字列とすることにより、テキスト中に存在した場合には（語の境界が不明瞭であるため）タグ付けを見送った文字列を、新規語として辞書に登録することができる。

## 5.3. NE タガー改良

本研究の手法は、NE タガーの改良手法として有効である。初期のタガーに十分なデータやルールがない場合にも、この手法を用いて実験・評価することにより、性能向上のために追加すべきデータやルールを明らかにすることができる。

## 5.4. 今後の方針と課題

今回の実験で用いた手法は再帰的に利用することができるので、本来は、用語リストのうちどれくらいの用語に最終的に属性がつけられるか、を確認して、属性がつかずに残ったもの分析して原因を検討する必要がある。これについては継続的に実験を行い、稿を改めて報告したい。

また、今回の属性拡張ルールは少しシンプルに過ぎるくらいがあるので、形態素解析結果の情報などを用いて条件判断を少し複雑にすることを考えたい。

課題としては、形態素解析レベルで分割に失敗している場合には、属性ラベル・属性影響語とも信頼度が低くなってしまふことがあげられる。このような場合には、用語を獲得する際に用いた統計情報を用いるか、または、（コストを高くしたくないという当初の要求からは反するが）原文における周囲の語との関係をチェックすることが必要となると考えられる。

## 6. まとめ

情報抽出をベースとしたアプリケーションに利用される（固有表現）タガー用の辞書作成のために、Web ページなどの大規模文書集合から統計的に獲得された用語に属性ラベルを付与することを試みた。ただし、応用アプリケーションでの継続運用を想定しているので、処理にかかるコストを抑えることに留意した。

与えられた用語リストを、まず基本的な属性拡張ルールを適用することにより、属性ラベルが付与できるもの3グループと未付与の用語1グループに分割した。さらに、すでに属性ラベルが付与された用語のうち、専門用語と判定されたものの構成要素を属性影響語と考え、うち特定の属性を持っていない語に専門用語属性を仮設定して、未付与グループの用語リストに再度属性拡張ルールを適用した。

実験の結果、本手法により、獲得用語の 3/4 程度に対して半自動で属性ラベルを付与できる可能性があることがわかった。

## 文 献

- [1] 中村達生, 産学連携支援ツール (Bluesilk<sup>®</sup>) の仕組み, 情報管理, Vol. 46, No. 7, pp.455-462, Oct.2003.
- [2] 産学連携支援ツール Bluesilk<sup>®</sup>, <http://www.bluesilk.biz/>
- [3] 濱口佳孝, 池野篤司, 井佐原均, “Web からの情報抽出・検索システムにおける全文検索,” 情報処理学会研究会報告, 情報学基礎研究会, Vol.76, No.2, Sept.2004.
- [4] 山本英子, 池野篤司, 濱口佳孝, 井佐原均, “検索支援に向けた Web 文書集合からの用語獲得,” 情報処理学会研究報告, 自然言語処理研究会, Vol.164, No.30, (電子情報通信学会言語理解とコミュニケーション研究会), Nov.2004.
- [5] 形態素解析システム 茶筌, <http://chasen.aist-nara.ac.jp/index.html.ja>
- [6] 汎用連想検索エンジン GETA, <http://geta.ex.nii.ac.jp/>
- [7] 福本淳一, 下畑光夫, 榎井文人, “固有名詞抽出における日本語と英語の比較,” 情報処理学会研究会報告, 自然言語処理研究会, Vol.126, No.15, Jul.1998.
- [8] 大沼宏行, 池野篤司, “ホームページやメールを対象とした質問応答システム,” 情報アクセスのためのテキスト処理シンポジウム発表論文集, pp.89-95, Feb.2003.