Vol.2012-ICS-166 No.1
2012/3/11

Web ニュース記事データを用いた 見出し文の意味的具体化

稲井聡† 芋野美紗子† 土屋誠司†† 渡部広一†

人と共存するロボットを開発するためには，まず人とロボット間のコミュニケーション手段に会話インターフェイスを導入する事が望まれている．人とロボット間の会話に Web ニュースサイトから抽出した見出し文を組み込むことによって，人はロボットにより親しみを感じるようになる．しかしながら，見出し文原文では内容が抽象的であるため，会話に応用することが出来ない．
よって，上記の問題点を解決するため，本稿では概念ベースと関連度計算を用いた見出し文の意味的具体化手法を提案する．

Concrete meaning generation of the headline Using the data for Web news articles

Satoshi Inai† , Misako Imono†, Seiji Tsuchiya†† and Hirokazu Watabe†

Developing a robot to coexist with people, people hope that adopting a conversational interface to a way of communication between people. By incorporating headlines collected from Web-News sites into a conversation between people and robot, people think that robots are more familiar to them. However, the original news headlines can’t apply into the conversation between people and robot because it’s abstract depiction.
To present the concrete headline, this study proposes Concrete Meaning Generation Method of headline with a Concept-Base and a Calculation Method of Degree of Association.

1. はじめに

情報技術の発展・少子高齢化による労働人口の減少などにより，今後人のパートナーとして自律的に行動できる知的ロボットの実現が望まれている．そこで問題になってくるのが，人とロボットのインターフェースである．従来，産業用ロボットなどを操作していたリモコン・キーボードのようなインターフェースでは，操作方法取得の訓練が必要であり，ユーザにとって大きな負担になる．よって，今後ロボットが一般人のパートナーとして活躍するためにも，人の会話を理解した上で行動できる知的ロボットの実用化が重要になってくると考えられる．

現在，人と円滑にコミュニケーションが行える知的ロボットの実現に向けた研究が幅広く行われており，人とロボットが円滑な会話を行うために，人からの質問などに応答するシステムなどについても関心が高まっている．しかし，人からの質問に応答するシステムだけでは，人がロボットに対して一方的に話し掛けるのみであり，人とロボットの相互会話を発展させることは不可能である．そこで人とロボットが双方向でコミュニケーションを行える様にするため，人からの質問に応答するシステムの他に，ロボットからユーザに話題を提供するシステムの研究・開発も盛んに行われている．そのため，人とロボットの会話を発展させ，かつ人に最新の話題を提供する上で，Web ニュースサイトなどの時事情報の収集が重要になってくると考えられる．

人の興味を引く話題を提供するために，各新聞社の Web サイトから収集したニュース見出し文の知識ベース(以降，本稿では時事情報知識ベースと称す)を用いて，嗜好・会話などに関する様々な研究[1][2]が行われている．しかし，各新聞社に掲載されている Web ニュース見出し文はその掲載スペースの関係上，「オリンパス元社長，社長復帰断念」のように，時刻や場所・動詞などが省略されている場合が多く，また表現も抽象的である．よって，ニュース見出し文をそのまま会話文に使用すると不自然になってしまうという問題点が存在する．従って，ユーザに話題提供できる時事情報を収集するため，そのニュース見出し文の欠けている情報を自動で補い，具体化する必要があると考えられる．

本稿では，新聞社の Web サイトから収集したニュース見出し文の構造解析を行った上で，欠けている情報の追加・補完を行う技術の提案を行う．この技術により，ユーザにより具体性のある時事情報を提供する事が可能になる．

† 同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University
†† 同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

2. 研究概要

本論文ではユーザに，より具体的な時事情報知識を提供できるシステムを提案する．本システムでは以下の処理を行うものとする．

1)各新聞社の Web サイトから収集したニュース見出し文の構造解析を行い，欠けている情報(動詞・場所，時刻など)が何か調べる．
2)その見出し文に対応するニュース記事本文から抽出した自立語群（以降，本稿ではニュース記事データと称す）から，ニュース見出し文の欠けている情報を取得・補完する．
3)補完されたニュース見出し文を時事情報知識ベースに格納する．

この上記の処理により，見出し文「オリンパス元社長，社長職復帰断念」を「マイケル・ウッドフォード元オリンパス社長が 5 日夜社長職復帰を断念する.」というような具体性のある時事情報に変換することが可能となる．図 1 は見出し文「オリンパス元社長，社長職復帰断念」に対する処理を表したものである．

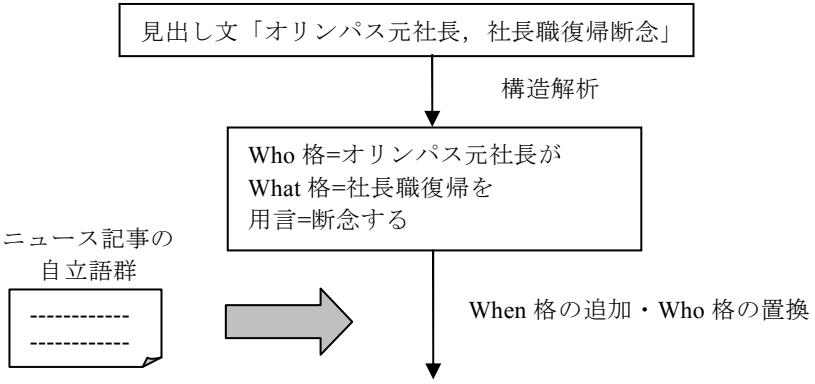

図 1　本システムの処理

3. 使用技術

3.1 概念ベースと関連度計算方式

概念ベース[3]は，語（概念）の特徴を表す語（属性）を大量に集めたものであり，属性には重みが定義されている．本研究では，複数の国語辞書や新聞などから抽出した概念や属性を加えた約 12 万の概念からなる概念ベースを使用する．
図 2 に概念ベースの構造を示す．

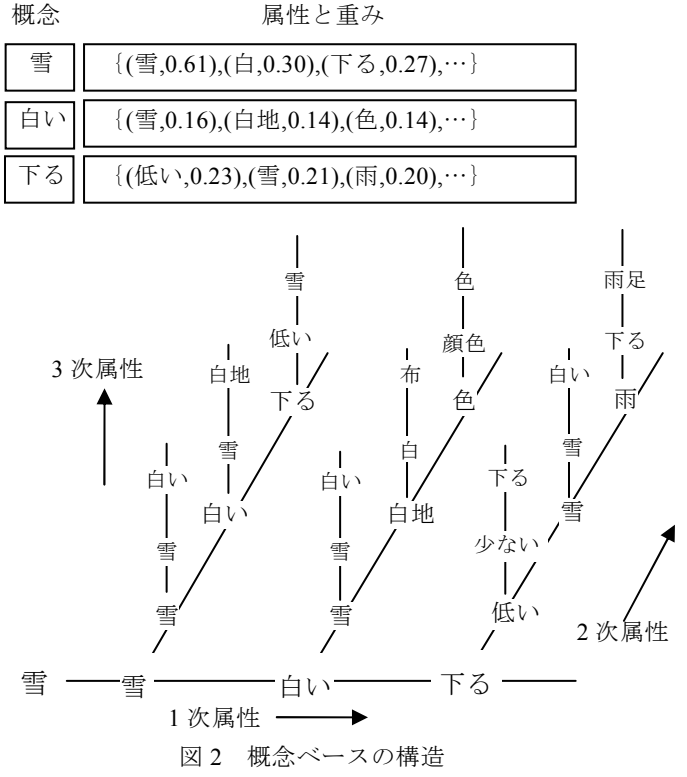

図 2　概念ベースの構造

また，概念ベースに登録されていない語のことを未定義語と呼ぶ．

関連度計算方式[4]とは，概念と概念の関連の強さを 0.0 から 1.0 までの値で定量的に評価するものである．各概念を 2 次属性まで展開し，重みを考慮した属性集合の一致度合いを計算する．表 1 に関連度計算の具体例を示す．

表 1　関連度計算の例

基準概念	対象概念	関連度
自動車	車	0.919
	自転車	0.343
	猫	0.003

3.2 NTT シソーラス

NTT シソーラス[5]とは単語の意味や概念を分類，整理して用語を階層的に体系化したものである．各節点に相当する語をノード，ノードに含まれる語をリーフと呼ぶ．

NTT シソーラスには，一般名詞の意味的用法を表したものと用言の文型パターンを示したものがある．前者は，一般名詞の意味的用法を表す約 2700 個の意味属性（ノード）の上位下位関係・全体部分関係が木構造で示されたものであり，約 13 万語（リーフ）が登録されている．後者は，日本語用言約 6000 語に対し，その用言がとる文型パターンを示したものもある．本稿で用いているのは前者のほうである．

NTT シソーラスの一部を図 3 に示す．

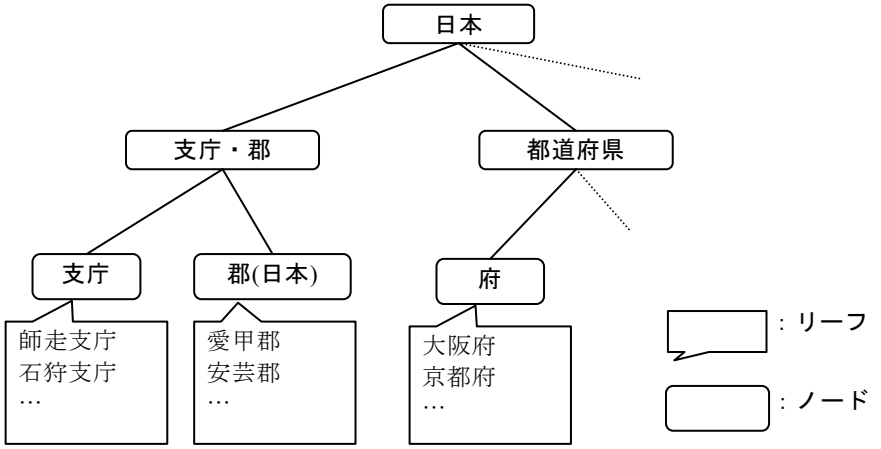

図 3　NTT シソーラスの一部

3.3 TF・IDF

TF・IDF 法[6]とは，語の頻度と網羅性に基づいた重み付け手法である．*TF* はある文書中 *d* に出現する索引語 *t*（文書の内容を構成する要素）の頻度を表す尺度である．*IDF* はある索引語が全文書中のどれくらいの文書に出現するか（特定性）を表す尺度であり，式(3.1)で定義される．なお，*N* が検索対象となる文書集合中の全文書数，*df*(*t*) が索引語 *t* が出現する文書数である．

$$idf(t) = \log_e \frac{N}{df(t)} + 1 \qquad (3.1)$$

3.4 Web-IDF

Web-IDF は Web にある文書のみを用いて索引語の出現頻度を考慮する手法である．Web-IDF では式(3.1)の *N* を Google が保有している日本語のページ数，*df*(*t*)を索引語 *t* の Google で検索を行ったときのヒット件数とする．なお，Google は全言語において保有しているページ数は公開されているが，日本語のページとして保有している数は公開されていないため，日本語の文書として最も使われている主格の助詞「は」で検索を行ったヒット件数（13,620,000,000 件-2012 年 1 月現在）を Google が保有している日本語の全ページ数としている．

3.5 未定義語の属性獲得手法

未定義語の属性獲得手法[7]とは，未定義語 *X* の意味的特徴を表す属性（単語）とその重要性を表す重みの組を Web を用いて獲得する手法である．

以下の 1)から 4)までが，その手法の流れである．

1) 未定義語 *X* をロボット型検索エンジン[8]に入力し，検索結果ページを獲得する．
2) 獲得した検索結果ページに対して形態素解析を行い，自立語を獲得する．
3) 獲得した検索結果ページに含まれる自立語の出現頻度と Web-IDF の算出を行い，TF・Web-IDF 重み付けを行う．
4) 自立語を重み順に並び替え，なおかつ，概念ベースに存在する自立語とその重みを *X* の属性として抽出する．

この手法を用いて未定義語 *X* の属性とその重みの組を構成する．未定義語 *X* の属性は式 3.2 のように構成される．なお，式(3.2)の x_i は *X* の一次属性，w_i はその属性に対する重みである．

$$X = \{(x_1, w_1), (x_2, w_2) \cdots, (x_n, w_n)\} \qquad (3.2)$$

この作業により，未定義語に属性が与えられるため，未定義語に対しても関連度を算出することが可能となる．本稿では，この未定義語の属性獲得手法をオートフィードバック（*Auto Feedback*：*AF*）と呼ぶ．

3.6 Web から構築した大規模格フレーム

Web から自動構築した大規模格フレーム[9]とは，動詞とその動詞に関係する名詞を用法ごとに整理したものである．この格フレームを用いることにより，動詞からその動詞に結びつく名詞，格，頻度などのデータを取得できる．表 2 は格フレームに名詞「鉛筆」を，表 3 は名詞「鉛筆」と動詞「削る」を入力した結果である．

表 2　名詞「鉛筆」における格フレームの出力

動詞	格	頻度
描く	デ格	371
書く	デ格	321
削る	ヲ格	217
はしる	ヲ格	201
・・・	・・・	・・・

表 3　名詞「鉛筆」と動詞「削る」における格フレームの出力

格	頻度
デ格	426
ノ格	5
ニ格	2
ガ格	1

この大規模格フレームにより，名詞「鉛筆」と動詞「削る」の間に入る格「デ格」が他の格「ノ格」「ニ格」などと比べ，最も頻度が高いということが表 3 で分かる．

4. 提案システムの流れ

本稿で提案する手法の流れは以下の通りである．(図 4) まず，各新聞社の Web サイトから収集したニュース見出し文の構造解析を行い，欠けている情報(動詞・場所，時刻など)が何か調べる．次に，その見出し文に対応するニュース記事の自立語群から，ニュース見出し文の欠けている情報を取得・補完する．最後に補完されたニュース見出し文を時事情報知識ベースに格納する．

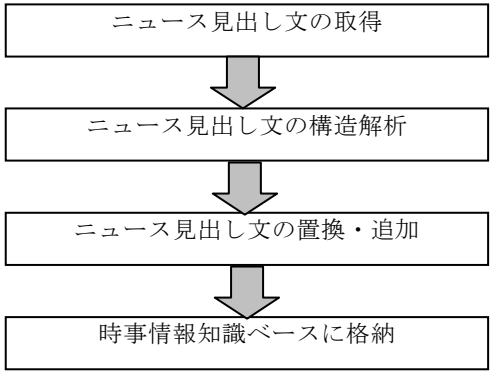

図 4　提案システムの流れ

4.1 ニュース見出し文の取得

本システムでは,「YOMIURI ONLINE（読売新聞）[10]」・「asahi.com（朝日新聞）[11]」・「毎日 jp（毎日新聞）[12]」の 3 社の新聞社の Web サイトに表示されているニュース見出し文を使用する．なお，今回の研究では記者特有の視点で書かれたコラムや社説などに対応した見出し文は取り扱わない．

4.2 ニュース見出し文の構造解析

Web から獲得したニュース見出し文は掲載スペースの関係上，見出し文「イスラエルがガザ空爆」のようにサ変名詞「空爆」の後ろに動詞「する」が省略されるなどといったケースが多いため，会話文として不自然である．そこで，コンピュータにニュース見出し文の主語・述語などを正しく理解させるため，見出し文を 6W1H(Who, What, When, Where, Whom, Why, How)と用言およびニュースの内容を表す Theme に分類する．この処理により，コンピュータは主語である Who 格と用言などを理解した上で，見出し文「イスラエルがガザ空爆」を主語・述語などが明確な文（以降，本稿では自然文と称す）「イスラエルがガザ空爆する.」に変換することが可能になる．

ニュース見出し文の構造解析の一例を図 5 に示す．

まず, 見出し文「イスラエルがガザ空爆　パレスチナ民兵 1 人死亡」に全角空白「　」が含まれているので，見出し文を 2 文に分割する．次に，分割された見出し文の単文の係り受け解析を実施する．この処理により，文節「イスラエルが」が文節「ガザ空爆」に，文節「パレスチナ民兵 1 人」が文節「死亡」に係っていることが分かる．
この時，各単文の最後の文節「ガザ空爆」「死亡」が形態素解析により，「空爆」「死亡」がサ変名詞であることが分かるため，「イスラエルがガザ空爆」の用言に「ガザ空爆する」を，「パレスチナ民兵 1 人死亡」の用言に「死亡する」を格納することが出来る．

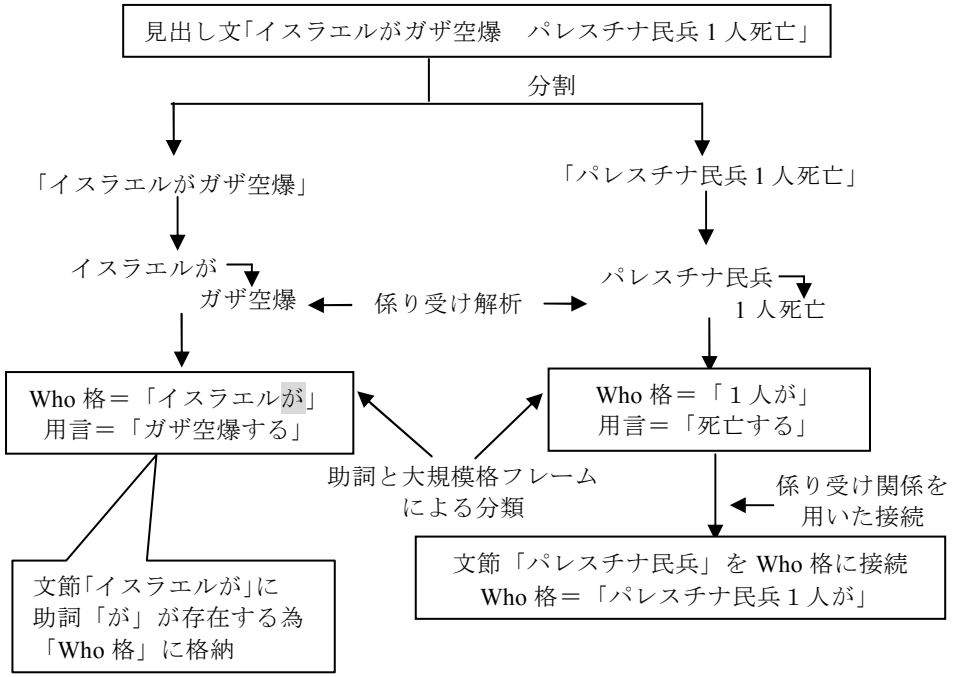

図 5　見出し文の構造解析の一例

次に，助詞と大規模格フレームによる処理を行う．「イスラエルがガザ空爆」の場合，文節「イスラエルが」に主語を表す助詞「が」が含まれているので，「イスラエルが」が「イスラエルがガザ空爆」の「Who 格」に格納される．そして，「パレスチナ民兵 1 人死亡」の場合，文節「1 人死亡」の名詞「1 人」と「死亡」の間に助詞「が」が入ることが大規模格フレームで分かるので，「パレスチナ民兵 1 人死亡」の「Who 格」に「1 人が」を格納することが出来る．

最後に係り受け関係を用いた分類の処理を行う．「パレスチナ民兵 1 人死亡」の文節「パレスチナ民兵」が「1 人死亡」に係っていることが，係り受け解析で分かっているので，「パレスチナ民兵 1 人死亡」の Who 格「1 人が」の前に，文節「パレスチナ民兵」を接続する．この処理により，「パレスチナ民兵 1 人死亡」の Who 格が「パレスチナ民兵 1 人が」に変換される．

以上の見出し文の構造解析により，全ての文節を 6W1H+用言+Theme に分類することが可能になる．

4.3　ニュース見出し文の置換・追加

ニュース記事本文から抽出した自立語群であるニュース記事データを用いて，見出し文に時刻・場所・用言の追加・Who 格の置換の処理を行う．図 6 に見出し文「オリンパス元社長，社長職復帰断念」の記事本文から抽出した自立語群（ニュース記事データ）の一例を示す．

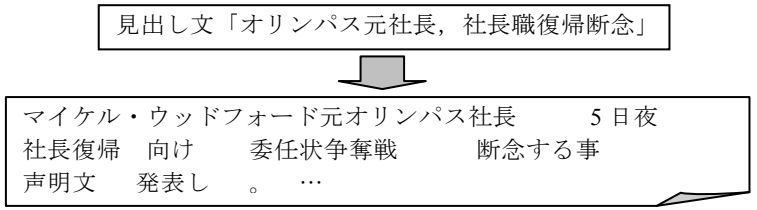

図 6　ニュース記事データの一例

図 7 に，見出し文「オリンパス元社長，社長職復帰断念」に対する Who 格の置換・時刻（When 格）の追加の処理の流れを示す．

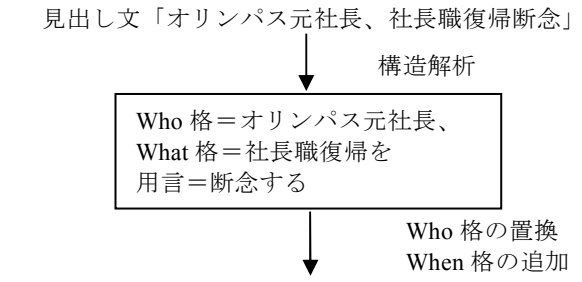

図 7　Who 格の置換・When 格の追加の一例

図 7 では，見出し文の構造解析を行った後，「オリンパス元社長、社長職復帰断念」のニュース記事データから，When 格として追加可能な「時間」に関係する語句「5 日夜」を NTT シソーラスで見つけ出し，見出し文に追加する．次に，Who 格と部分一致する「マイケル・ウッドフォード元オリンパス社長」を取得した後，「Who 格+用言」と「マイケル・ウッドフォード元オリンパス社長+用言」の共起ヒット件数を調べる(表４)．

表 4　共起ヒット件数

入力文	共起ヒット件数
オリンパス元社長、断念する	188000
マイケル･ウッドフォード元オリンパス社長が断念する	305, 000

最後に「Who 格」と 「マイケル・ウッドフォード元オリンパス社長」の関連度を調べる．この時，関連度が 0.1 以上かつ 2 つの共起ヒット件数に大きな誤差が見当たらなかったため，Who 格と「マイケル・ウッドフォード元オリンパス社長」の置換を実施する．

以上の見出し文に対する置換・追加処理により，見出し文「オリンパス元社長、社長職復帰断念」を「マイケル・ウッドフォード元オリンパス社長が 5 日夜社長職復帰を断念する.」というような具体性のある時事情報に変換することが可能になる．

5. 評価

新聞社の Web サイトから獲得した見出し文 120 文をシステムにかけ，その出力文を被験者 3 人に見てもらい，評価を行った．表 5 は評価で使用した見出し文とその出力文である．

表 5　見出し文と出力文

入力文	出力文	評価
小沢元代表、維新の会「方向性は同じだ」 ： 民主党	維新の会に関して民主党の小沢一郎元代表が、３０日「方向性は同じだ」と発言する．	○ ○ ○
民主・斎藤恭紀議員、離党表明…追随の動きも ： 政治 ：	民主党の斎藤恭紀衆院議員が、２７日午前離党を表明する．追随の動きも挙げる．	○ ○ ×
オウム・平田信容疑者、逮捕監禁致死容疑で逮捕 ： 社会 ：	平田信(まこと)容疑者（４６）が、１日朝警視庁(でorに)逮捕監禁致死容疑で逮捕する．	× × ×

この評価実験により，全体で 58.3%の確率で文法的に正しいかつ具体性のある時事情報を出力することが出来た．

6. おわりに

本稿では，ニュース見出し文とその記事本文から自立語だけを抽出したニュース記事データと関連度計算などを用いて，具体性のある時事情報を提供するシステムを提案した．

結果として，58.3%の精度でユーザに見出し文の内容を具体化した時事情報を提供することが可能になった．今後の展望として，Whom 格・What 格の置換の機能などを追加することで，より具体性のある時事情報をユーザに提供出来ると考えられる．

7. 謝辞

本研究の一部は，科学研究費補助金(若手研究(B)21700241)の補助を受けて行った．

参考文献

1) 河合智弘，吉村枝里子，土屋誠司，渡部広一，“個人情報に基づく時事情報提供システムの構築”，電子情報通信学会技術研究報告, Vol.109, No.439, pp23-28, 2010
2) 吉岡孝治，吉村枝里子，土屋誠司，渡部広一，“常識的連想によるニュースヘッドラインからの会話文生成”，情報処理学会研究報告. 2010-ICS-158 No.4, 2010
3) 奥村紀之，北川晋也，渡部広一，河岡司，“概念ベースの分析と精錬”，同志社大学理工学研究報告，Vol.46，No.3，pp.133-141，2005.
4) 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.
5) NTT コミュニケーション科学研究所監修，”日本語語彙体系”，岩波書店，1997.
6) 徳永健伸，“言語処理と計算 5 情報検索と言語処理”，東京大学出版会，1999.
7) 辻泰希，渡部広一，河岡司，“www を用いた概念ベースにない新概念およびその属性獲得手法”，第 18 回人工知能学会全国大会論文集，2D1-01，2003.
8) “Google”，http://www.google.co.jp/
9) 河原大輔，黒橋禎夫，“高機能計算環境を用いた Web からの大規模格フレーム構築”，情報処理学会自然言語処理研究会資料，2006-NL-171-12，pp.67-73，2006.
10) “ニュース　速報　YOMIURI ONLINE（読売新聞）”，http://www.yomiuri.co.jp/
11) “asahi.com：朝日新聞社の速報ニュースサイト”，http://www.asahi.com/
12) “毎日ｊｐ - 毎日新聞のニュース・情報サイト”，http://www.mainichi.jp/