2-2 大量の自然言語テキストへの情報アクセス技術

2-2 Information Access Technologies for Processing a Very Large Number of Natural Language Documents

村田真樹
MURATA Masaki

要旨

情報検索、情報抽出（テキストマイニング）、質問応答処理、文書自動分類など、自然言語情報に対する種々の情報アクセス技術を開発した。これらの技術の有効性は、評価型ワークショップ NTCIR において数多く 1 位の精度を出したことで確認されている。電子的文書の数は日々増加しており、これらの情報アクセス技術は一層重要なものとなってきている。

We have developed various information access technologies for information retrieval, information extraction (text mining), question answering, and document classification used in processing natural language documents. The effectiveness of these technologies was confirmed when they produced the highest level of precision in the NTCIR evaluation workshop. As the number of electronic documents continues to increase, these information access technologies will become increasingly useful.

[キーワード]
情報検索，情報抽出，テキストマイニング，質問応答処理，文書自動分類
Information retrieval, Information extraction, Text mining, Question answering, Document classification

1 まえがき

電子的文書の数は日々増加しており、これらの情報アクセス技術は一層重要なものとなってきている。この背景を考慮して、我々は、情報検索、情報抽出、質問応答処理、文書自動分類など、自然言語情報に対する種々の情報アクセス技術を開発した。これらの技術の有効性は、評価型ワークショップ NTCIR において数多く 1 位の精度を出したことで確認されている。本稿ではこれらについて記述する。

2 情報検索

情報検索とは、キーワード又は文で記述したユーザの検索したい内容を含む文書を、大量の文書群から取り出す技術のことである。特に、近年の自然言語処理では、キーワードよりも、むしろ、文で記述したユーザの検索したい内容を含む文書を取り出すことが主になっている[1][2]。

例えば、図 1 のようにユーザの検索したい内容を文章で記述する。この内容に合致する文書を検索する。例えば、図 2 のような文書（記事）を検索する。

一般に情報検索では、図 1 のユーザの検索したい内容を形態素解析（単語に分割し各単語の品詞を求める）をして、名詞の単語を取り出し、その単語（キーワードと呼ばれる）をより多く含む文書を検索する。また、めったに出現しない希少価値の高い単語の重みを大きくし、どの文書にでも出現する単語を希少価値の低い単語として重みを小さくする。そして、重みの大きい単語をより多く含む文書を検索する。具体的な単語の重み付けの方法には幾つかあるが、我々は BM25 と呼ばれる

重み付け法[3]を利用した。この方法は、高性能の重み付け法として知られている。情報検索には、tf・idf 法と呼ばれる方法がある。tf・idf 法では、検索対象の文書で出現するキーワードの個数を tf、そのキーワードが出現するデータベース中の文書の数を df、文書総数を N として、tf・log(N/df) をキーワードの重みとして計算する方法である。BM25は tf・idf 法を改良したもので、tf・idf 法の tf の影響を弱めた式を利用する。

また、擬似関連性フィードバック法と呼ばれる方法[4]も利用した。これは、一回目の文書検索を行い、その文書検索の結果の上位の記事群に偏って多く出現している単語を取り出す。そして、それらの単語を検索のキーワードに加えて、二回目の文書検索を行う。検索に用いるキーワードでは、上位の記事群に偏って多く出現している単語については、単語の重みが大きくなるような重み付けを行う。そのように単語の重み付けをして、二回目の文書検索を行い、その結果を最終的な文書検索の結果とする方法である。一回目の文書検索を行った結果、その文書検索の結果の上位の記事群に偏って多く出現している単語は、元のキーワードの類義語であることが多く、これらキーワードも追加して利用することで、文書検索の性能をあげる方法である。この方法は有効であることが知られている。我々はこの方法も利用した。

上述の BM25 と擬似関連性フィードバック法に加えて、我々は、文書のタイトルの情報を利用する方法を用いた。文書のタイトルに検索のキーワードが出現した場合、特にその文書は、検索内容に合致する文書である可能性が高いと考えた。そして、重みの大きいキーワードがなるべく文書のタイトルに出現している文書を検索する方法を考案した。この方法を用いることで、高性能な情報検索を実現する技術を構築した[1][2]。

情報検索の評価型ワークショップ NTCIR と呼ばれるものがある[5]。ここでは、同じ問題を複数の団体で解き、精度を比較する。情報検索の精度評価では、検索して取ってくるべき文書を、検索結果においてどれだけ上位で多く取れるかを調べる。我々はこのワークショップに参加して数多く 1 位の精度をあげた[1][2][6]。このことから、我々の技術が高いことが証明されている。

テーマ：企業合併
説明：記事には企業合併成立の発表が述べられており、その合併に参加する企業の名前が認定できる事。また、合併企業の分野、目的など具体的内容のいずれかが認定できる事。企業合併は企業併合、企業統合、企業買収も含む。

図1 ユーザの検索したい内容

キグナス石油精製を東燃が 100 %子会社化

東燃は十六日、系列のキグナス石油精製（資本金十億円、本社・川崎市、森利英社長）を一〇〇%子会社化すると発表した。同社は東燃が七割、ニチモウが三割出資しており、東燃はニチモウが所有する全株式六十万株を百二十五億円で買収する。

東燃は石油精製専業大手で、設備シェアは一九九三年度末で八%。キグナス石油精製を加えると九・四%にシェアがアップする。

ニチモウは山口県の工場閉鎖などに伴う経費ねん出のため株式譲渡を決断した。

石油業界は来年春の特定石油製品輸入暫定措置法（特石法）廃止をにらみ、コスト削減と効率化を進めており、グループ企業統合を含む再編の動きがいよいよ本格化してきた。

図2 検索された記事

3 情報抽出（テキストマイニング）

情報抽出とは、大量の自然言語データから有用な情報を抽出する技術のことである。テキストマイニングとも呼ばれる。情報抽出として、文献書誌情報などを分析して、研究動向を調査する研究を行った。ここでは、言語処理学会誌「自然言語処理」を対象に行った結果を紹介する[7][8]。

テキストマイニングの基礎は、時系列データなどで、単語の出現頻度などを数えて、どういった

単語が時系列的に増加したか減少したかを分析し、興味がどのように移り変わったかを調べることである。

自然言語処理学会誌に掲載された論文において、論文のタイトルに出現した単語を取り出した。その単語を、研究分野を示すキーワードとしてとらえて、そのキーワードをタイトルに含む論文の頻度を時系列的に調査した。その結果を図 3 に示す。ただし、研究分野としてふさわしくないキーワード（例：「的」「研究」）は人手で取り除いた。また、学会の第 1 巻（1 年次）から 10 巻（10 年次）までを対象とした。およそ 1994 年から 2003 年までのものである。

図 3 では等高線表示を利用した。等高線の高さ（色の濃さ）が件数を示す。各研究分野の発表件数のデータにおいて、発表される年次の平均値と最頻値と中央値を求めてそれらの平均を求め、図ではこの平均の値の小さい順に上から下に表示した。各研究分野を示す単語には合計件数と上述の平均の値を付記した。このため、図では早い年次に偏って発表件数の多い分野は上の方に、遅い年次に偏って発表件数の多い分野は下の方に表示される。この表示は、早い年次又は遅い年次に偏っている研究分野を容易に認識できるという効果を持つ。この表示方法は、テキストマイニングのいろいろな場面で役立つと思われる。この表示技術は現在特許出願中である。

図 3 から「日本語」「解析」が特に多いことが分かる。論文誌では、「動詞」「名詞」「解消」「確率」「コーパス」「多義」などが図の上の方に現れ、これらの研究分野が早い年次に盛んであったことが分かる。「形態素」「係り」「対話」「音声」は 6 年次に盛んであること、「要約」「検索」「翻訳」などが遅い年次で盛んになったことも分かる。特に「要約」は 6 年次と 9 年次でそれの特集号が出たためそのときに偏って多く出現している。「翻訳」は増加傾向にあることがうかがえ、今後も発表件数が増加することが予測される。

次に、研究組織についても調査を行った。研究発表した論文を数えた。その結果を図 4 に示す。

この調査でも各分野の発表件数のデータにおいて、発表される年次の平均値と最頻値と中央値を求めてそれらの平均を求め、図ではこの平均の値の小さい順に表示した。各研究組織には合計件数と上述の平均の値を付記した。図では早い年次に偏って発表件数の多い組織は上の方に、遅い年次に偏って発表件数の多い組織は下の方に表示される。ここでは合計件数の多かった組織のみを表示した。組織名が変わった組織については頻度の最も大きかった名称を利用して表示している。

これらの図から通信総合研究所（現在の情報通信研究機構）、ATR（現在、言語処理グループを中心に情報通信研究機構に所属）の発表件数が多いことが分かる。自然言語処理の研究分野において、我々の研究組織（情報通信研究機構）は卓越した成果をあげていたことが分かる。また、NTT、ATR は早い年次から多くの発表をしているが、通信総研と東京大学は 10 年の年次の中では比較

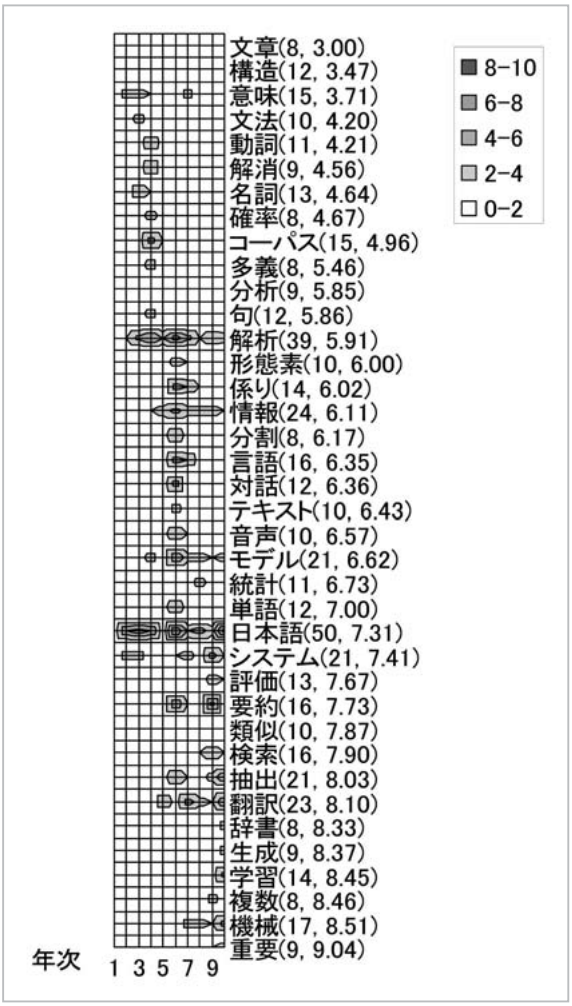


図3　分野ごとの発表件数の推移

等高線の高さ（色の濃さ）が件数を示す。図中の各分野を示す単語に付与した二つの数字は左が合計件数を右が発表件数の多い年次の平均を意味する（厳密な定義は本文を参照のこと）。

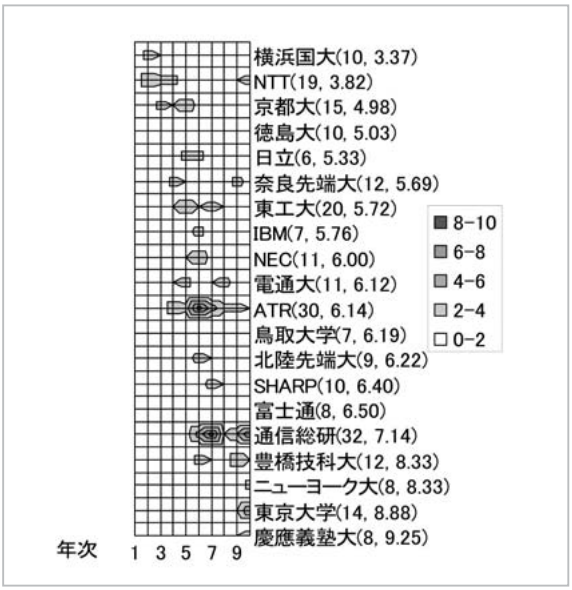


図4 研究組織ごとの発表件数の推移

等高線の高さ(色の濃さ)が件数を示す。図中の各研究組織に付与した二つの数字は左が合計件数を右が発表件数の多い年次の平均を意味する(厳密な定義は本文を参照のこと)。

的後ろの方の年次で多くの発表をしていることが分かる。通信総研と東京大学は図から増加傾向にあることが読み取れ、今後も発表件数が増加することが予測される。その他の組織についてもどの年次で多く発表しているかは、この図を参照することで容易に知ることができる。

4 質問応答システム

質問応答システムとは、例えば人間が「富士山の高さは何メートルですか」という質問をすると、「3776 メートル」と的確にその質問の解答を答えるシステムのことである(図 5)。あらかじめ質問と答えの知識に関するデータベースを整備するのではなく、大量の自然言語で書かれた文章から解答を取り出して出力するところに特徴がある。自動的に計算機が質問に対して答えを答えるため、あたかもその計算機に知能があるかごとくみえる場合もある。質問応答システムは、計算機が人間と同じように思考できるようにする人工知能の研究につながる、興味深い研究テーマである[9]。

図5 質問応答処理の様子

4.1 質問応答システムの重要性

質問応答システムは主に以下の三つの観点から重要である。

(1) 情報検索(文書検索)よりも便利

質問応答システムはユーザの質問に対して答えそのものを提示するため、ユーザは文書を読む必要もなく自分の知りたい情報を容易に取得することができる。情報検索システムでは関連文書を提示するだけなので、ユーザは欲しい情報を見つけるために提示された関連文書をいちいち読まなければならない。

(2) 膨大な自然言語データからの知識抽出

自然言語で書かれた膨大なデータから自由自在に知識を取り出すことができると便利である。質問応答システムは、この知識の取り出しを、自然言語による質問に対して解答を出すという形で実現することになる。

(3) 他の知識処理システムの構成要素としての利用

自然言語による質問に対して解答を出すという質問応答システムは、他の知識処理システムの構成要素として利用できる可能性がある。例えば、指示表現の指示先の推定においても、「塩と水をまぜます。その食塩水を…」といった文において「その食塩水」が指すものを推定するときに、「塩と水をまぜると何ができますか」と質問応答システムに問い合わせ「食塩水」と解答を得て、それを指示先と推定するということができる。このような利用は、指示先の推定の問題に限らず、様々な自然言語処理システム、知識処理システムで可能であり、質問応答システムは他の知的処理システムの一構成要素として利用できる可能性がある。

質問応答システムは人工知能システムを作成する場合には最低限必要な技術になると思われ、将来の知的処理・知識処理の根幹システムになると思われる。

4.2 質問応答システムの一般的構成

現在の質問応答システムの一般的な構成は以下のとおりである。

(1) 解表現の推定

システムは疑問代名詞の表現などに基づいて解表現（解がどのような言語表現か）を推定する。例えば、入力の質問文が「日本の面積はどのくらいですか」だとすると「どのくらい」という表現から解表現は数値表現であろうと推測する。

(2) 文書検索

システムは質問文からキーワードを取り出し、これらのキーワードを用いて文書を検索する。この検索により、解が書いてありそうな文書群を集めることになる。例えば、入力の質問文が「日本の面積はどのくらいですか」だとすると、「日本」「面積」がキーワードとして抽出され、これらを含む文書を検索することになる。

(3) 解の抽出

システムは解が書いてありそうな文書群から、推定した解表現に適合する言語表現を抽出し、それを解として出力する。例えば、入力の質問文が「日本の面積はどのくらいですか」だとすると、文書検索で検索した「日本」「面積」を含む文書群から、解表現として推定した数値表現にあたる言語表現を解として抽出する。

4.3 我々の質問応答システム

我々の質問応答システムも前節の方法を採用している。我々の質問応答システムでは、さらに、解の抽出の際に、推定した解表現に適合する言語表現のうち、質問文から取り出したキーワードのなるべく近くの表現を解として抽出する方法を利用している。さらに、複数の記事群の情報を総合的に利用することで簡便に性能をあげる方法も考案して用いている。しばしば質問の解は複数の記事で発見される。そのような場合は一つの記事に基づくより複数の記事を使って解の推定を行った方がより良い解を得ることができる。そこで、我々の方法では、複数の記事から得られた解の候補の得点を加算することで複数の記事の情報を利用する。しかし、単純に得点を加算したのでは逆にシステムの性能を下げる場合がある。そこで、我々の方法では、この単純に加算する欠点を減らすために、得点を少しずつ減らしながら加算する。我々の質問応答システムは、これらの方法を利用する[10]。

情報検索の評価型ワークショップ NTCIR では、質問応答のタスクもある。ここでは、質問応答処理に関する同じ問題を複数の団体で解き、精度を比較する。我々はこのワークショップに参加して数多く 1 位の精度をあげている[11][12]。このことから、我々の技術が高いことが証明されている。

5 文書自動分類

ここでの文書の自動分類は、文書をあらかじめ決められた分類に自動分類する技術を指す。文書の自動分類は、大量の文書を分類、整理するときに役立つ、情報アクセスにおいて重要な技術である。我々は、文書の自動分類として、特許文書の自動分類の研究を行った（図 6）。

特許文書には F タームと呼ばれる分類が付与されている。一つの特許に複数の F タームが付与される。F タームとは、特許分類のために日本で独自に用いられているもので、目的や用途、構

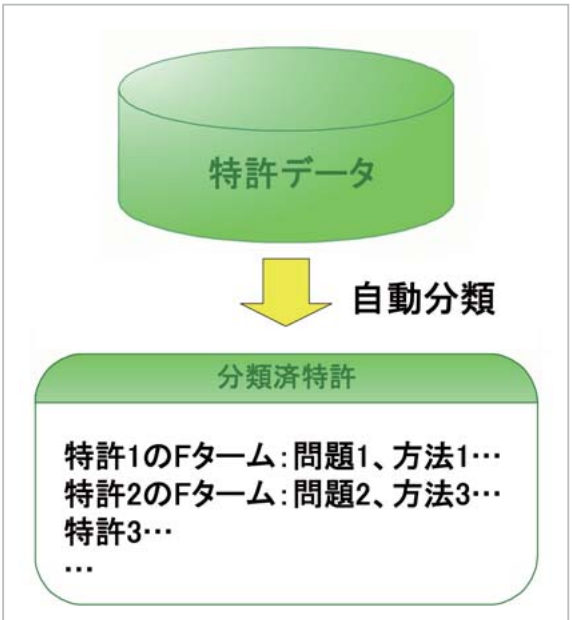


図6 特許の自動分類

造や材料など、様々な観点から特許を見るのに適した分類体系となっている。F タームを利用することで、従来よりも特許の特徴を細かく分析することが可能となる。例えば、問題と方法の観点を表にすると、どのような特許取得の可能性が残されているかを判断することができる（図 7）。

各特許に付与すべき F タームを推定する研究を行った。我々は k 近傍法を改良した方法を利用した。k 近傍法[13]とは、分類を付与すべき事例と最も類似する k 個の事例を集めて、それらの事例で最も多く付与されている分類をその付与すべき事例の分類とする方法である。我々はこれを改良し、k 個の最も類似する事例の中でも、分類を付与すべき事例との類似度が高い事例に付与されている分類ほど解になりやすいという工夫を k 近傍法に追加した。それら解になりやすい分類のうちどれまでを解とするかを自動で学習する枠組みを追加した。また、k 個の類似する事例を集める際には、事例同士の類似度を求める必要性があるが、これに情報検索で性能が高いとされている BM25 などの方法を利用した。これらの方法を用いて F タームの自動付与を行った。

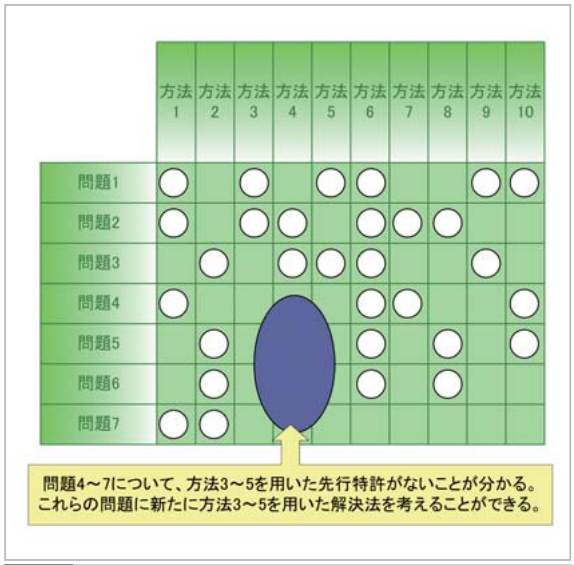


図7 新しい特許の可能性の発見

情報検索の評価型ワークショップ NTCIR5 では、特許文書の F ターム分類のタスクもあった。ここでは、特許文書の F ターム分類に関する同じ問題を複数の団体で解き、精度を比較する。我々はこのワークショップに参加して 1 位の精度をあげた[14]。我々の手法は、上述のように改良型の k 近傍法を用いた。他の参加チームは、単語ベクトルを用いたベクトル空間モデルや、機械学習のサポートベクトルマシン法を用いていた。我々は、他のチームに比べて、15 %程度、高精度な結果を得た。

6 むすび

情報検索、情報抽出、質問応答処理、文書自動分類など、自然言語情報に対する種々の情報アクセス技術について述べた。電子的文書の数は日々増加しており、これらの情報アクセス技術は一層重要なものとなってきている。

参考文献

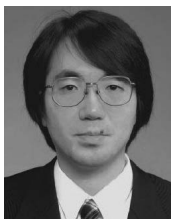
1 村田真樹，馬　青，内元清貴，小作浩美，内山将夫，井佐原均，“位置情報と分野情報を用いた情報検索”，自然言語処理（言語処理学会誌），７巻，２号，p.141-160，2000.

2 Masaki Murata, Qing Ma, Kiyotaka Uchimoto, Hiromi Ozaku, Masao Utiyama and Hitoshi Isahara, "Japanese Probabilistic Information Retrieval Using Location and Category Information", IRAL'2000, Hong Kong, Sep.30, 2000.

3 Stephen E. Robertson, Steve Walker, Susan Jones, Micheline. Hancock-Beaulieu, and Mike Gatford, "Okapi at TREC-3", Proceedings of the third Text REtrieval Conference (TREC-3), p.109-126, 1994.

4 Lisa Ballesteros and W. Bruce Croft, "Phrasal translation and query expansion techniques for cross-language information retrieval", In Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '97), p.84-91, 1997.

5 小川泰嗣，佐々木裕，増山　繁，村田真樹，吉岡真治，“参加者から見た　NTCIR”，人工知能学会，17巻，3号，p.306-311，2002.
6 Masaki Murata, Qing Ma, and Hitoshi Isahara, "Applying Multiple Characteristics and Techniques to Obtain High Levels of Performance in Information Retrieval", Proceedings of the NTCIR Workshop 3 (CLIR), 2002.
7 村田真樹，一井康二，馬　青，白土　保，金丸敏幸，井佐原均，過去10年間の言語処理学会論文誌・年次大会発表における研究動向調査，言語処理学会ホームページ（http://www.nak.ics.keio.ac.jp/NLP/trend-survey.html），2007.
8 Masaki Murata, Koji Ichii, Qing Ma, Tamotsu Shirado, Toshiyuki Kanamaru and Hitoshi Isahara, "Trend Survey on Japanese Natural Language Processing Studies over the Last Decade", The Second International Joint Conference on Natural Language Processing, Companion Volume to the Proceedings of Conference including Posters/Demos and Tutorial Abstracts, p.252-257, Jeju Island, Korea, Oct. 2005.
9 村田真樹，“質問応答システムの現状と展望”，電子情報通信学会，86巻，12号，p.959-963，2003.
10 Masaki Murata, Masao Utiyama, and Hitoshi Isahara, "Use of Multiple Documents as Evidence with Decreased Adding in a Japanese Question-answering System", Journal of Natural Language Processing, Vol.12, No.2, p.209-247, 2005.
11 村田真樹，内山将夫，白土　保，井佐原均，“シリーズ型質問文に対して単純結合法を利用した逓減的加点質問応答システム”，システム制御情報学会論文誌，20巻，8号，2007.
12 Masaki Murata, Masao Utiyama and Hitoshi Isahara, "Japanese Question-Answering System For Contextual Questions Using Simple Connection Method, Decreased Adding with Multiple Answers, and Selection by Ratio", Asia Information Retrieval Symposium (AIRS) 2006, Shangri-La's Rasa Sentosa Resort, Singapore, Oct. 16, p.601-607, 2006.
13 Yiming Yang and Xiu Liu, "A re-examination of text categorization methods", Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.
14 Masaki Murata, Toshiyuki Kanamaru, Tamotsu Shirado, and Hitoshi Isahara, "Automatic F-term Classification of Japanese Patent Documents Using the k-Nearest Neighborhood Method and the SMART Weighting", Journal of Natural Language Processing, Vol.14, No.1, p.163-190, 2007.



村田真樹（むらたまさき）
知識創成コミュニケーション研究センター自然言語グループ主任研究員（旧情報通信部門けいはんな情報通信融合研究センター自然言語グループ主任研究員）　博士（工学）
自然言語処理