多言語情報源を対象とした意味的連想検索実現のためのメタデータ自動翻訳方式

An Automatic Metadata Translation Method for Semantic Associative Search in Multilingual Information Resources

大橋 英博† 清木 康‡
石黒 晶子§

Hidehiro OHASHI Yasushi KIYOKI
Akiko ISHIGURO

本稿では,翻訳辞書群を用いた信頼度依存の多数決処理によって,専門分野のメタデータを自動翻訳する方式を示す.本方式は,複数の翻訳辞書に信頼度を設定し,信頼度を反映させた多数決処理により翻訳語を決定することにより,翻訳精度を向上させる方式である.また,本方式を多言語情報源を対象とした意味的連想検索に適用する方式を示す.本方式を意味的連想検索に適用することにより,意味空間の言語的制約を取り除くことができ,専門家の知識の利用範囲を大幅に拡大させることが可能となる.

In this paper, we present an automatic metadata translation method based on majority decision reflecting reliability of several translation dictionaries. In this method, we set reliability to several translation dictionaries, and select a translated word through the majority decision process. By using this method, we can improve precision of translation. We also present an application of the automatic metadata translation method to semantic associative search. We apply this method to extend the scope for using the semantic space which is originally created depending on the specific language. By this extension, we obtain a semantic associative search environment supporting multiple language for representing queries and retrieval-candidate information resources.

1. はじめに

本研究の目的は,多言語環境において,意味的連想検索[5][6]を実現することである.そのために多言語情報源を対象としたメタデータ自動翻訳方式を示す.ここでは,メタデータとは検索対象ドキュメントの内容を言葉によって表現したデータであり,検索時に問い合わせとの比較対象として用いるものである.

大量の文書情報を格納したドキュメントデータベースから情報を検索する手法として,パターンマッチングによる検索手法が普及している.しかし,検索者の求める情報を取得するためには,パターンマッチングによる検索手法では不十分である.大量のドキュメントの中から検索者にとって価値のある情報を見つけ出すためには,検索者の文脈を解釈する機構を有する検索手法が必要となる.検索者の文脈を解釈する機構を有する検索手法として,意味的連想検索方式[5][6]が提案されている.この方式における文脈とは, 検索語のもつ多義性を解消するために検索語と共に与えるものであり,検索語の意味を確定するための情報である.この方式は,専門家の知識を意味空間として体系化し,検索者が入力した検索キーワードと各ドキュメントとの意味的な類似性を計量する空間として使用する点に特徴がある.ここで,専門家の知識（専門知識）とは,意味空間を形成するための特徴付きベクトルの作成に用いる専門分野辞書および事典に記述されている当該分野に関する情報をさす.意味的連想検索方式により,検索者の文脈を専門家の知識体系に基づいて解釈し,検索キーワードと各ドキュメントとの意味的な類似性に基づいて,ドキュメント検索を行うことが可能となる.

意味的連想検索方式は専門知識を意味空間として体系化する.ここで意味空間とは,意味の数学モデル[5][6]に基づいて構成されたn次元正規直交ベクトル空間をさす.意味の数学モデルでは,当該分野の専門用語と,専門用語を説明するための単語（特徴語）を使用して専門用語の特徴付きベクトルを作成する.特徴付きベクトルは各専門用語ごとに作成する.特徴付きベクトル群から意味空間生成用マトリクスを作成する.意味空間生成用マトリクスを直交化し, n次元正規直交ベクトル空間（意味空間）を作成する.意味の数学モデルでは,専門知識を表現するために,当該専門分野の専門用語と,専門用語を説明する特徴語をある一つの言語を用いて記述する.この意味空間を用いた意味的連想検索を行う時には,専門用語および特徴語を記述したときに言語を用いなければならない.ある言語で記述された意味空間を別の言語環境で利用するためには,意味空間を当該言語を用いて再作成する必要がある.意味空間は,特定分野の辞書や事典を用いて,専門家の手により手動で作成されることが一般的である.そのため,意味空間の作成には時間と労力を要する.意味空間は専門家の知識を数学的手法[5][6]を用いて体系化した高度な計量空間である.この空間は本来特定の言語環境によらず,あらゆる言語環境のドキュメント検索に適用できる計量空間である.

本研究では,意味的連想検索方式[5][6]において使用するメタデータを,他の言語に自動翻訳する方式を提案する. 提案方式を使用することで,意味空間を再作成することなく,多言語環境において意味的連想検索を実現することができる.これにより,専門家の知識の利用範囲を拡大させることが可能となる.

2. メタデータ自動翻訳方式

メタデータ自動翻訳方式を次に示す.提案方式では,次の手順でメタデータの翻訳を行う.

(1)メタデータ翻訳に用いる辞書の選択および信頼度設定
(2)対訳表作成
(3)対訳表を用いたメタデータ自動翻訳

各手順の詳細を説明する.

2.1 メタデータ翻訳に用いる辞書の選択方法および信頼度設定方式

メタデータの翻訳を行うための翻訳辞書を選択する.ここ

† 学生会員 慶應義塾大学大学院政策・メディア研究科
hohashi@sfc.keio.ac.jp
‡ 正会員 慶應義塾大学環境情報学部
kiyoki@sfc.keio.ac.jp
§ 株式会社レクサー・リサーチ ishiguro@lexer.co.jp

で,意味空間を記述するために使用した言語を L_m,検索キーワードおよび検索対象ドキュメントを記述している言語を L_lとする.以下の方針に基づいて,L_mから L_lへの翻訳を行う3種類の辞書を選択する.

- $Dict_s$: 当該分野において最も専門性の高い翻訳辞書
- $Dict_b$: 当該分野において基本的な翻訳辞書
- $Dict_g$: 分野によらない一般的な翻訳辞書

ここで,$Dict_s$は当該分野の専門用語を翻訳するために使用する辞書である.$Dict_b$は当該専門分野の基本語を翻訳するための辞書である.$Dict_g$は分野によらない一般的な単語を翻訳するための辞書である.一般に,単語の収録数は多い順に $Dict_g$, $Dict_b$, $Dict_s$の順となる.

これらの $Dict_s$, $Dict_b$, $Dict_g$を使用してメタデータの翻訳を行う.ここで,3 辞書の信頼度設定を行う.信頼度設定はその辞書が当該分野においてどの程度の専門性をもつかによって行う.ここでは翻訳辞書 D の信頼度 r(D)を以下のように定義する.

$$r(D)=\begin{cases}3 & D=Dict_s\ のとき\\ 2 & D=Dict_b\ のとき\\ 1 & D=Dict_g\ のとき\end{cases}$$

当該分野において,$Dict_s$は専門性の高い単語を翻訳できる辞書であるため,最も高い信頼度 $r(Dict_s)=3$ を設定する.$Dict_b$は,当該分野において,基本的な単語を翻訳できる辞書であるため,$r(Dict_b)=2$ を設定する. $Dict_g$は分野によらない一般的な単語を翻訳する辞書であるため,$r(Dict_g)=1$ を設定する.

2.2 対訳表作成方式

提案方式では,翻訳を行うための対訳表を作成する.対訳表は表2.2に示す形式となっている.対訳表は次の手順で作成する.

言語1(L_l)	言語m(L_m)	優先度	辞書情報ビット列

表2.2 対訳表の形式

Table 2.2 Translation Table Format

(1) 翻訳辞書群の信頼度を用いて優先度計算表を作成する.優先度計算表は表2.3 に示す形式によって構成する.優先度は次のように計算する.

優先度	参照辞書	辞書情報ビット列 DictBits		
7	A,B,C	1	1	1
6	A,B	1	1	0
5	A,C	1	0	1
4	A	1	0	0
3	B,C	0	1	1
2	B	0	1	0
1	C	0	0	1

表2.3 優先度計算表

Table 2.3 Priority Calculation Table

※ただし,辞書の信頼度は 辞書A＞辞書B＞辞書C とする.

(a) 翻訳辞書ビット列 DictBits

辞書の信頼度の値により,各辞書を各ビット列に対応させる.信頼度 n の辞書を第 n ビットに対応させる.例えば,信頼度3の辞書は第3ビットに対応する.このビット列を翻訳辞書ビット列 DictBits と呼ぶことにする.

(b) w_m と w_l とが互いに対訳であるとき,$w_m \Leftrightarrow w_l$ とする.翻訳辞書ビット列を設定する.翻訳元言語 L_m の中の単語 w_m を各辞書で調べ,L_lの単語 w_lを得たとする.このとき, w_m に対する翻訳語 w_lの翻訳辞書ビット列 DictBits は以下のように設定する.DictBitsの第 n ビットの値 bit(n)を次のように定義する.ただし,辞書 D において w_m が w_lの訳であることを $(w_m \Leftrightarrow w_l)\in D$ と表す.

$$bit(n)=\begin{cases}1 & (w_m\Leftrightarrow w_l)\in D \cap r(D)=n\\ 0 & それ以外のとき\end{cases}$$

この bit(n)によって DictBits の各ビットの値を設定する.

(c) 翻訳辞書ビット列 DictBits を 10 進数に変換した値を対訳 $w_m \Leftrightarrow w_l$ の優先度 $p(w_m \Leftrightarrow w_l)$とする.

優先度 $p(w_m \Leftrightarrow w_l)$は次のように定義する.

$p(w_m \Leftrightarrow w_l) = priority(DictBits, w_m, w_l)$

ここで,priority(DictBits, w_m, w_l)は, w_m と w_l,および優先度計算表によって定まる DictBits(2進数)を,10進数に変換する関数とする.優先度計算表を表 2.3 に示す.

(2) 言語 L_m のメタデータ群 $w_{mi}(i=1,...,n)$を用意する.この w_{mi} を対訳表の L_m の単語群として設定する.

(3) 翻訳辞書群から翻訳元言語 L_mの中の単語 w_mの訳である言語 L_lの単語 w_lを調べ,対訳表中の w_lの欄に設定する.

(4) w_mの訳である w_lを参照した辞書から,辞書情報ビット列 DictBits を作成する.例えば w_mに対する訳 w_lが $Dict_s$の辞書から得られたとする.このとき,$w_m \Leftrightarrow w_l$ の DictBits の $r(Dict_s)$列に 1 を立てる.

(5) 辞書情報ビット列 DictBits から優先度 $p(w_m \Leftrightarrow w_l)$を計算する

上記の(1)～(5)の手順により,対訳表を作ることができる.表 2.4 にこの手順によって作成された対訳表の例を示す.

日本語	英語	優先度	辞書情報ビット列
アルコール依存症	Alcohol dependency	1	001
アルコール中毒	Alcoholism	6	110
…	…	…	…
イタイイタイ病	Itai-Itai disease	4	100

表 2.4 対訳表

Table 2.4 Translation Table

2.3 対訳表を用いたメタデータ自動翻訳方式

対訳表の優先度を用いたメタデータの自動翻訳方式を示す.メタデータの翻訳においては,言語 L_l を翻訳元言語とし,言語 L_mを翻訳先言語とする.対訳表中の各エントリーには翻訳元言語 L_lの単語 w_lと,翻訳先言語 L_mの単語 wm の対訳と翻訳辞書ビット列,および優先度が設定されている.この優先度の最も大きい訳を使用して,メタデータの自動翻訳を行う.w_lの訳 w_{sm}を次のように決定する.

(1) 対訳表の中で,言語 L_lの中の単語 w_lに対応する言語 L_m中の単語が n 個存在したとする.この単語群を w_{mi} (i=1,...,n)とする.

(2) $w_l \Leftrightarrow w_{mi}$の翻訳に関する優先度 p は次のように表す.

$p(w_l \Leftrightarrow w_{mi})$ (i=1,...,n)

(3) w_{mi}(i=1,...,n)の中で優先度 $p(w_l \Leftrightarrow w_{mi})$が最大となる単語 w_{mi}を w_lに対する訳 w_{sm}とする.

$w_{sm} := w_{mi}$ in $\max(p(w_l \Leftrightarrow w_{mi}))$ (i=1,...,n)

2.4 意味的連想検索方式への適用

提案方式を意味的連想検索方式[5][6]に適用する方法を示す.意味的連想検索方式では,当該分野の専門用語と専門用語を説明する特徴語を使って,意味空間(n 次元正規直交ベクトル空間)を構成するための特徴付きベクトルを作成する.専門用語は検索対象ドキュメントに付与するメタデータ,および検索キーワードとして使用される.この専門用語の翻訳に提案方式を適用する.これにより,意味的連想検索方式で使用さ

れる基本データ,検索対象ドキュメントに付与するメタデータ,検索キーワードの自動翻訳が可能となる.

3. 提案方式を適用した多言語情報源を対象とした意味的連想検索システム

提案方式を適用した意味的連想検索システムの実現例を示す.この例では, 医療分野を対象とした既存の日本語の意味空間上において,同分野の英語の新聞記事に対する意味的連想検索を行なうシステムを実現した.図 3.1 にシステム構成図を示す.

3.1 システム構成図

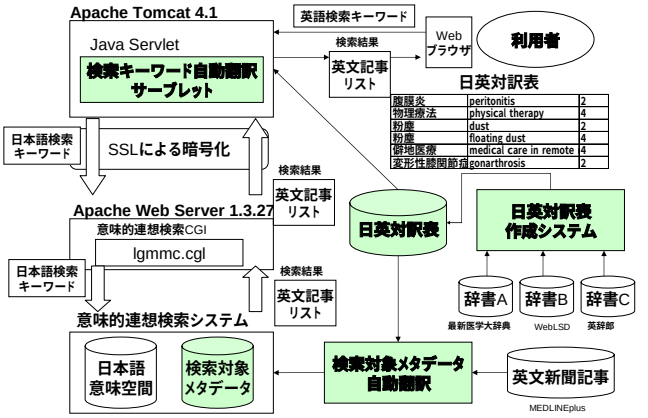

図 3.1 多言語情報源を対象とした意味的連想検索システム
Fig.3.1 Semantic Associative Search System in Multilingual Information Resources

3.2 対訳表作成機能

対訳表の作成には,医学大辞典[2], WebLSD[3], 英辞郎[1]の 3 辞書を使用した.ここでは,医学大辞典を医療分野の専門辞書($Dict_s$)とした.WebLSD を同分野の基本辞書($Dict_b$)とした.英辞郎を分野によらない一般的な翻訳辞書($Dict_g$)とした.

3.3 検索対象メタデータ自動翻訳機能

本方式では,次の３ステップで検索対象メタデータの自動翻訳を行う.

(1) 対訳表の言語 L_l 欄にある単語（英単語）の集合を W_l とする.

(2) 英文記事 a_i の中の単語 $w_a \in a_i$ を調べ, $w_l \in W_l$ を含んでいた場合,a_i のメタデータとして w_l を割当てる.

(3) w_l を言語 L_m（日本語）の単語 w_m に翻訳する.

本方式を適用し,対訳表の中で w_l に対応する w_{mi} (i=1,..,n)の中で優先度 $p(w_l \Leftrightarrow w_{mi})$(i=1,..,n)が大きい単語を w_l に対する訳 w_m とする.

この方法を全ての英文記事に対して行う.これにより,各英文記事に対して,日本語の検索対象メタデータが付与される.本実現システムでは,検索対象ドキュメントとして,MEDLINEplus[4]に収録された医療に関する英文で書かれた新聞記事 578 件を使用した.このデータを基にして,上記の方法で検索対象メタデータを抽出し,自動翻訳を行った.

3.4 検索キーワード自動翻訳機能

この実現例では日本語意味空間を用いた意味検索を行うため,言語 L_l（英語）で表現された検索キーワード k_l を,言語 L_m（日本語）に翻訳する必要がある.英語の検索キーワード k_l を,次のように日本語検索キーワード k_{sm} に自動翻訳する.

(1) 検索キーワード k_l は対訳表の言語 L_l（英語）の単語群 W_l に含まれる単語とする.

(2) 対訳表の中に k_l に対応する言語 L_m（日本語）の単語が n 個存在したとする.この単語群を w_{mi}(i=1,...,n)とする.この w_{mi} (i=1,...,n)の中で,最も優先度の大きい単語 w_{sm} を k_l に対する訳 k_{sm} とする.

4. 実験

4.1 実験の概要

この実験では,本方式によるメタデータの自動翻訳実験を実施した.あらかじめ人手により準備したメタデータの正解データと,本方式によって作成された翻訳結果とを比較して,翻訳精度を検証した.

4.2 実験に使用した翻訳辞書

実験には,医学大辞典[2], WebLSD [3], 英辞郎 [1]の 3 辞書を使用した.同様に,医学大辞典を医療分野の専門辞書($Dict_s$)とした.WebLSD を医療分野の基本辞書($Dict_b$)とした.英辞郎を分野によらない一般的な翻訳辞書($Dict_g$)とした.これらの辞書を使用して,医療分野における対訳表を作成した.対訳表の言語 L_m（日本語）の単語群として,医学意味空間の基本語群 691 語を使用した.この結果,2506 エントリーの医学分野の対訳表を作成した.

4.3 正解データ

医療に関する 100 個の英単語を用意し,それらに対して人手により日本語訳を割り振った.この 100 個の単語の組を正解データとして使用した.正解データ例を表 4.3 に示す.

No.	英語	日本語
1	AD	アトピー性皮膚炎
2	AIDS	エイズ
…	…	…
100	Alcoholism	アルコール中毒

表 4.3 正解データの例
Table 4.3 The Examples of Correct Data

4.4 翻訳精度の計算方法

自動翻訳によって得られた単語のうち,正解データと一致した単語の数を,正解データ数(=100)で割った値を翻訳精度とした.

c: 正解データと一致した単語数
n: 全正解単語数(=100), 翻訳精度 p = c / n * 100

4.5 実験方法

次の５項目について実験を行った.

(1) 実験 1：分野によらない一般翻訳辞書($Dict_g$)のみによる自動翻訳
(2) 実験 2：当該分野の基本辞書($Dict_b$)のみによる自動翻訳
(3) 実験 3：当該分野の専門辞書($Dict_s$)のみによる自動翻訳
(4) 実験 4：上記の 3 辞書を使用した多数決による自動翻訳

分野によらない一般翻訳辞書,当該分野の基本辞書,および当該分野の専門辞書の 3 辞書を使用して対訳表を作成し,自動翻訳を行った.この実験では,3 辞書の多数決によって訳の優先度を決定し,その優先度の最も大きい訳を選択して自動翻訳を行った.実験 4 の優先度は,当該対訳を得ることができた辞書の数を設定した.

(5) 実験 5：本方式-信頼度依存の多数決による自動翻訳

3 辞書に信頼度を設定し,その信頼度と多数決によって優先度 p を決定し,優先度 p が最も大きい訳を選択して自動翻訳を行った.

4.6 実験結果 1-翻訳精度の比較

実験結果を図 4.6 に示す.この結果から,いずれの種類の辞書も単独で使用した場合では十分な翻訳精度が得られないことが明らかになった.3 辞書の多数決によって,翻訳精度は 80 と大幅に向上している.さらに本方式である辞書に信頼度を設定し,その上で多数決を行う方式では,翻訳精度は 85 と最も高い値を示している.

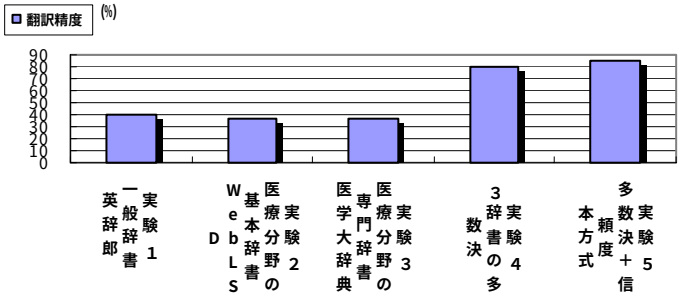

図 4.6 実験結果 翻訳精度の比較
Fig.4.6 Comparison of Translation Precision

4.7 実験結果 2-不正解データの内容の分析

次に各実験で不正解となった単語の内容について調べた.この結果を図 4.7 に示す.図 4.7 では不正解となった単語を,誤訳した単語と対訳表で発見できなかった単語とに分け,それぞれの全正解単語数に対する割合を「誤訳率」と「未発見率」として示している.この結果から,辞書の専門性が高くなるにつれ,誤訳率が低くなる一方で,未発見率が高くなっていることが明らかとなった.3 辞書の多数決により翻訳を行った場合では,未発見率は 0 になった.一方で誤訳率は 20 となり,分野によらない一般辞書の誤訳率と同じ値になっている.本方式では,未発見率は同様に 0 となっており,誤訳率が 15 となっている.本方式では,多数決処理のみを行う場合に比べて,誤訳率が減少している.

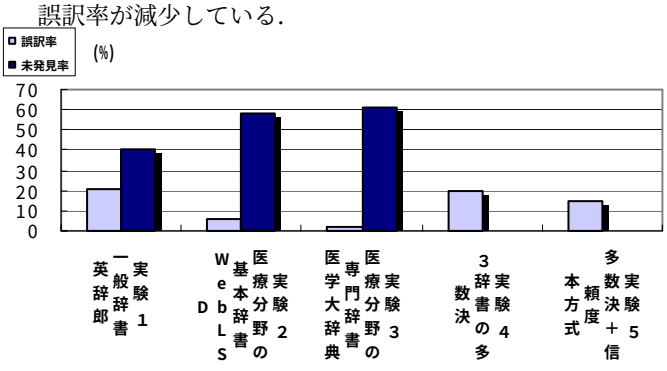

図 4.7 実験結果 不正解データの内容の分析
Fig.4.7 Analysis of Incorrect Data

5. 考察

実験結果より,それぞれの 1 翻訳辞書だけ使用した場合では,十分な翻訳精度が得られていない.3 辞書を使用した多数決による翻訳方法では,それぞれ 1 辞書だけを使用した場合より,翻訳精度が大きく向上している.本方式による翻訳により,多数決のみよる翻訳に比べてさらに翻訳精度が向上している.この結果から,本方式による自動翻訳の有効性が確認できた.

不正解データの内容を分析した結果から,複数の辞書を単純に合わせて使用するだけでは,誤訳率を上昇させてしまうことが明らかになった.複数の辞書を組み合わせて単語の未発見率を下げつつ,誤訳率の上昇を抑えるには,本方式による翻訳方式が有効である.例えば,対訳表の中には,英単語「pulsy」に対して日本語訳「脈」,「気分」が 2 語存在している.「pulsy ⇔ 脈」の優先度 p が 4(辞書情報ビット列 100),「pulsy ⇔ 気分」の優先度 p が 1 (辞書情報ビット列 001) となっている.正解を「pulsy ⇔ 脈」としたとき,本方式では,最も大きな優先度 p=4 を持つ「pulsy ⇔ 脈」が選択されるので必ず正解が得られる.多数決のみによる翻訳ではどちらの対訳も辞書情報ビット列中の"1"の数は 1 であるため,どちらも優先度 p が 1 となる.この場合,どちらの訳が適切か判断するための情報がないため,2 組の対訳の中から 1 組の対訳が非決定的に選択される.これらの理由から,多数決による翻訳に比べ,本方式は高い翻訳精度を実現できる.

6. まとめ

本稿では,専門分野におけるメタデータの自動翻訳方式を示した.本方式では信頼度を設定した複数の辞書の多数決により自動翻訳を行う.この方式により,複数の辞書の多数決のみによって自動翻訳を行う方法に比べて精度の高い自動翻訳を実現した.また,本方式を適用した多言語情報源を対象とした意味的連想検索システムを実現した.本方式を意味的連想検索システムに適用することで,意味空間の言語的制約を取り除き,専門家の知識の利用範囲を拡大することが可能となる.

[文献]

[1] 英辞郎,アルク,2002.
[2] 最新医学大辞典 CD-ROM 第 2 版,医歯薬出版, 1997.
[3] ライフサイエンス辞書 WebLSD , http://lsd.pharm.kyoto-u.ac.jp/WebLSD/ , 京都大学大学院薬学研究科医療薬理学分野, 2003.
[4] MEDLINEplus, http://www.nlm.nih.gov/medlineplus/, A service of the U.S. National Library of Medicine and the National Institutes of Health, 2003.
[5] Kiyoki, Y., Kitagawa, T. and Hayama, T.:``A metadatabase system for semantic image search by a mathematical model of meaning", ACM SIGMOD Record, vol. 23, no. 4, pp.34--41, 1994.
[6] Kiyoki, Y.,Kitagawa, T. and Hitomi, Y.: ``A fundamental framework for realizing semantic interoperability in a multidatabase environment", Journal of Integrated Computer-Aided Engineering, Vol.2, No.1, pp.3--20, John Wiley & Sons, Jan. 1995.
[7] 大橋 英博, 清木 康: ``情報通信分野を対象とした意味的連想検索機構によるWWW検索エンジンの実現,"情報処理学会研究報告, 2001-DBS-125(I), pp.233-240, 2001.

大橋 英博 Hidehiro OHASHI
慶應義塾大学大学院政策・メディア研究科修士課程. 2002 慶應義塾大学環境情報学部卒業.データベースシステムの研究に従事.

清木 康 Yasushi KIYOKI
慶應義塾大学環境情報学部教授. 1983 慶應義塾大学大学院工学研究科博士課程修了, 工学博士. データベースシステム, 知識ベースシステム, マルチメディアシステムの研究に従事. ACM, IEEE, 電子情報通信学会, 情報処理学会, 日本データベース学会各会員.

石黒 晶子 Akiko ISHIGURO
株式会社 レクサー・リサーチ
1998 東北大学大学院理学研究科博士課程修了, 理学博士.