様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　5　月 24 日現在

機関番号：１７１０４
研究種目：基盤研究（Ｃ）
研究期間：２００９　～　２０１１
課題番号：２１５００１４３
研究課題名（和文）　複数認識器の統合による音声及び画像の協調理解とマルチモーダル対話システムへの応用
研究課題名（英文）　Cooperative Understanding of Speeches and Images Using Multiple Recognizer and Its Application to Multimodal Dialogue System
研究代表者
遠藤　勉（ENDO　TSUTOMU）
九州工業大学・大学院情報工学研究院・教授
研究者番号：１０１１２２９４

研究成果の概要（和文）：マルチモーダル対話で人間の行う先ずマルチモダリティとして捉えるという処理の工学的実現を目指して，複数の認識器や特徴量を統合的に利用する各種の手法を提案し，言語と画像情報を用いた画像検索，複数認識器の統合による音声理解，バッチ型と逐次型の学習器を併用した手形状認識，言語及び非言語情報を利用した対話状況理解，多面的情報に基づく人物識別等を行うシステムの構築を通して，その有効性を確認した。

研究成果の概要（英文）：　We proposed a wide variety of methods to integrate several approaches and features for multimodal dialogue systems. We developed a Web based image retrieval system using linguistic and image features first. We also realized a multiple speech recognizer with hierarchical relations. For hand posture recognition, we combined online and offline machine learning techniques. We introduced context features and top-view images to person identification.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,000,000	300,000	1,300,000
2010 年度	600,000	180,000	780,000
2011 年度	500,000	150,000	650,000
年度	0	0	0
年度	0	0	0
総　計	2,100,000	630,000	2,730,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理，マルチモーダルインタフェース，音声理解，ジェスチャ認識

１．研究開始当初の背景

人間と機械との自然なインタフェースの実現を目指してマルチモーダル対話システムの開発が 1980 年代から続けられてきたが，多くのアプローチは，音声発話，ジェスチャ，視線，表情等のモダリティを独立の情報源と仮定して個々のモダリティを解析し，その結果を必要に応じて統合するという方式を採用している。その典型的な手順は，語，句，文といった分析単位が明確で，しかもコーパスに基づくアプローチの有用性が確認されている音声言語情報処理を先ず行い，照応表現や曖昧性を持つ表現等の限定された箇所についてのみジェスチャ等音声言語以外のモダリティの解析結果を随時追加していくというもので，一部の限定されたタスクでの応用例を除いて，必ずしも満足できる成果は得られていない。一方，人文系分野の研究者を中心に，人間はすべてのモダリティが統合された全体を先ず捉えた上で，状況に応じて

各モダリティでの表現に焦点を当てるという仮説の下，人間と人間，人間と物のマルチモーダルインタラクションのデータを収録して，これを用いたマルチモーダルコミュニケーションの分析が進められている。この情報を先ずマルチモダリティとして捉えるという人間の処理方式を，工学的にどのようにして実現するかが今後の重要な課題となる。

２．研究の目的

本研究は，対面コミュニケーション状況で観測されるマルチモダリティ(音声言語，表情，ジェスチャ，身体姿勢等が統合したもの)を複数の認識モジュール(音声認識，顔検出，人物識別，ジェスチャ認識等)で入力・解析し，タスクの特性に応じて適宜結果の統合を行いながら，言語情報と画像情報を協調理解する方式を開発するとともに，マルチモーダルインタフェースに応用して有効性を検証しようとするものである。

３．研究の方法

(1)コーパス作成ツールの開発

Web の文書を対象として，感情表現などのタグ付けを行うためのツールを作成した。事前に 5000 文に対して 7 種類の評価視点と 2 値の評価タグを 2 名のアノテータによってタグ付けした。タグ付け用の GUI を作成し，既存のタグ付きデータを利用して，効率的に新たな文章へタグ付けを行えるツールを開発した。ツールにはタグ付きデータを利用した関連語の強調機能，過去のデータを利用した類似事例の提示や他のアノテータによるタグ付け結果の提示などの機能を実装した。

(2)テキストと画像情報を利用した画像検索システム

「夏らしい画像」といった幅広い意味を持つ曖昧な言語表現に対して，Web から得られる関連語と画像間の類似度を統合的に利用した画像検索システムのプロトタイプを構築した。システムの外観を図 1 に示す。

検索システムは，まずユーザからクエリを受け付ける。そして，そのクエリを基に，Web の検索 API を利用して関連語を抽出する。関連語抽出では「<<クエリ>>といえば」というようなキーワードを適用する。関連語を用いて画像検索を行い，その結果をクラスタリングして，ユーザに提示する。画像処理では，Earth Mover Distance を用い，画像レベルの類似性を算出する．これにより，例えば「動物の」マウスと「PC の周辺機器の」マウスを分別できる。システムは対話的にユーザのクエリとフィードバックを受け取り，精度の高い画像検索システムを実現した。

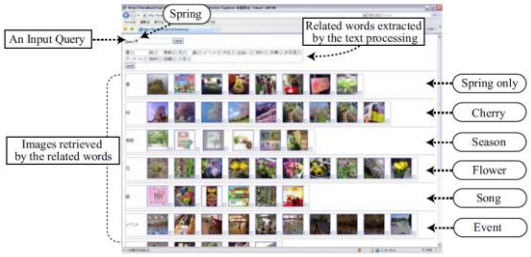

図 1：対話的画像検索システム

(3)複数の認識器を利用した音声理解手法

音声発話はマルチモーダル対話システムにおける重要な入力モダリティの一つである。音声による入力では，音声認識の正確さだけではなく，入力された音声が，そもそも対話システムに対する有効な情報なのかの判別が必要となる．入力音声のすべてがシステムへの命令である可能性は低く，むしろその多くは非命令（雑談）であることが多い。例えば，画像編集アプリケーションを想定し，そのシステムへの命令として「拡大」というコマンドがあるとする。このとき，「あれ？ちょっと拡大しすぎたかな」という発話に対して，キーワードスポッティング手法では「拡大」というキーワードを取り出し，誤って動作させてしまう可能性がある。このような命令ではない雑談に対しては，「命令ではない」と正しく棄却する機能が必須である。

そこで，命令のみを正しく受理するための小規模な音声認識器と雑談を検出するための大語彙認識器を併用し，命令発話に対してロバストで高精度な音声理解手法を実現した。具体的には複数の認識器の出力結果の編集距離を求め，最適な結果が何であるのかを推定した。さらに，多くの認識器を組み合わせる際，認識器の持つ特性や語彙に基づき階層化を行い，精度向上を図った。

(4)複数の認識器を利用した手形状認識とマルチモーダルインタフェース

ハンドジェスチャは，最も直感的な入力モダリティの一つである。本研究課題では，人物検出などのタスクでよく用いられる HOG 特徴（Histograms of Oriented Gradients）と機械学習による手形状認識を実現した。HOG を用いる手法では，一般に，Support Vector Machines（SVM）が用いられることが多い。しかしながら，一般に SVM には，十分な学習データが必要な問題や学習データに含まれていない人物での精度低下などの問題がある。そこで，SVM と逐次型の学習器を併用することで，ロバストでかつ個人に対して動的に最適なモデルを生成することのできる手法を提案した（図２参照）。

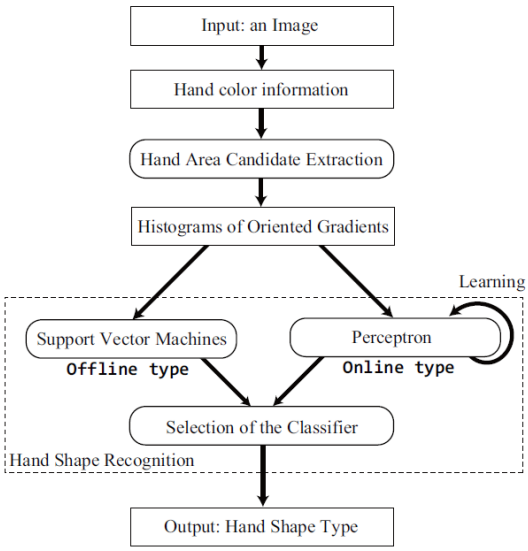

図 2：バッチ型学習器と逐次型学習器を統合した手形状認識手法

また，研究の方法(3)で示した音声理解手法と統合し，iTunes を操作するためのマルチモーダルインタフェースのプロトタイプを作成した。図 3 はその操作コマンドと動作例である。

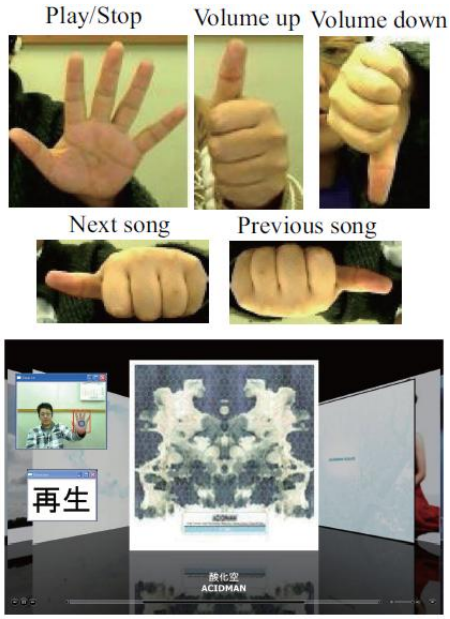

図 3：マルチモーダルインタフェース

(5) 言語と非言語情報を利用した対話の状態理解

円滑な対話を実現するためには，対話の状態を把握することが不可欠である。ここでは，複数人による対話の場面に対して，現在，その対話が盛り上がっているかどうかの判断を行う手法を提案した。盛り上がりの推定手法には，発話された言語的内容から推測する手法と声の高さなどの非言語情報を利用した手法があるが，人間の理解の過程を踏まえれば，図 4 に示すように，言語情報と非言語情報の両方を利用して推定することが自然である。

図 4：対話の盛り上がり判定

提案手法では，言語情報として，A1) 発話に出現する語彙，A2) 発話の長さ，A3) 発話における語彙の結束性を利用した。非言語情報としては，B1) 話者の声の高さ，B2) 発話の密集度，B3) 発話のタイミングの 3 つの特徴を利用した。これらの特徴量を発話から求め，ナイーブベイズ分類器に適用して盛り上がりの判定を行った。

(6) 画像を用いた人物識別

研究の方法(4)や(5)で示したように，誰がコンピュータの前で操作しているか，誰が発話をしているかなどの情報はシステム全体の精度向上には不可欠な要素である。そのためには精度の高い人物識別手法の確立が不可欠である。本研究課題では，2 つの点に着目して，研究を進めた。順にその詳細を示す。

①属性情報を利用した人物識別

人物を識別する手法としては，指紋，虹彩，音声など様々なものが存在するが，最も一般的で直感的なものは顔情報を利用した識別手法である。顔情報を利用した手法では，顔全体や目や口など顔の部品に対して，データベース中の画像との類似度を算出する手法が一般的である。しかしながら，このような顔情報のみ利用した手法では，例えば，顔の一部がマスクやメガネで隠れてしまった場合，精度が大きく低下する可能性が高い。

そこで本研究では，人物の属性情報を広い意味でその人物の持つ文脈情報と位置づけ，人物識別のタスクに適用する手法を提案した。具体的には，図 5 に示すように，画像中の人物が着ている衣服の特徴とその画像がいつ撮影されたかという時間特徴を人物識別の手法に導入した。衣服特徴では，大局的な面と局所的な面を考慮し，さらに，色とテクスチャの 2 つの点に着目して特徴量を抽出した。具体的には，色ヒストグラム（大局的）と色モザイク（局所的）およびパワースペクトル（大局的）と高次局所自己相関（局所的）である。これらの文脈特徴を顔情報と統合し，

AdaBoost 学習器に適用することで人物識別の精度向上を図った。

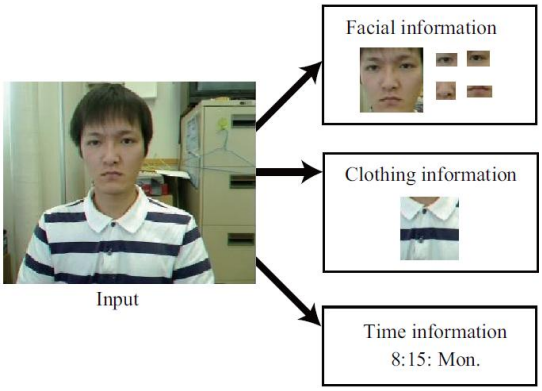

図 5：人物識別への文脈情報の適用

②頭上画像を利用した人物識別

研究の方法(6)-①では，顔の一部が隠れても，文脈特徴の利用によって，高い識別精度を実現している。しかしながら，状況によっては，人の重なりなどによって，顔だけでなく，衣服など様々な情報が獲得できない場合も少なくない。

そこで本研究では，頭上から撮影された画像に着目する。図 6 に示すように，頭上から撮影すれば，すべての隠れの問題が解消される。ここでは，ドアの前では人間は一度立ち止まるという仮定を置き，頭上方向から撮影された画像を用いて，人物を識別するタスクについて提案し，評価した。

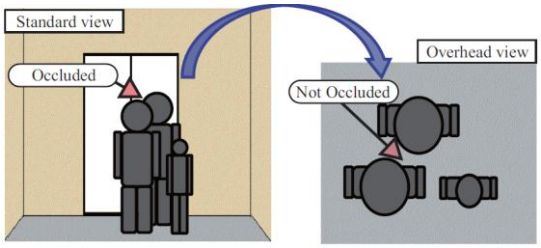

図 6：撮影環境による隠れの問題とその解決方法.

具体的には，まず背景差分法によって人物領域を検出する。続いて，獲得された人物領域画像から，図 7 に示すように，A)人物の体型特徴，B)髪の色特徴，C)髪型特徴，D)つむじ特徴の 4 つを抽出した。体型特徴では，人物領域の x 軸および y 軸の大きさを利用し，色情報としては輝度値を利用した。髪型とつむじ特徴では，まず，髪領域についてエッジ情報を抽出し，そのエッジ情報を基に，HOG 特徴量を適用した。つむじを取り出すためには，髪領域に一度平滑化フィルターを適用している。これらの特徴を AdaBoost 学習器に適用し，画像中の人物が誰なのかを識別した。

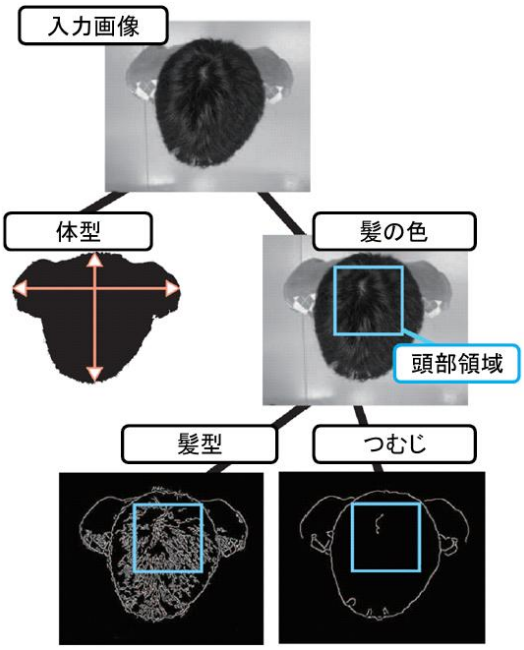

図 7：体型と髪特徴

４．研究成果

(1) コーパス作成ツールの開発

2 名の被験者により，2 回に分けてタグ付けに関する実験を行った。まず，350 文を用いて，実装した機能の有効性を検証した。その結果，評価表現の一致率が 5%程度，タグの κ 値が 0.2 程度上昇することを確認した。また，同じアノテータが，別の 450 文に対してタグ付けを行ったところ，一文あたりのタグ付けの作業時間が若干減少し，一方で κ 値が 0.1 程度上昇し，構築したコーパス作成ツールの有効性が確認された。

(2)テキストと画像情報を利用した画像検索システム

10 個のクエリについて評価したところ，単純な検索方法と比べて，提案手法では適合率が 30%弱向上することを確認した。また，上 n 件の検索結果の適合率を評価した場合でも，単純な手法では，n が大きくなるにつれ，適合率が大きく(n=1 から n=30 に変化した場合で 40%)低下したが，提案手法では，6%程度の精度低下しか生じておらず，提案手法の有効性が確認された。

(3)複数の認識器を利用した音声理解手法

148 個の発話例（内 88 発話が命令，50 発話が非命令）を用意し，6 名の被験者（男性：4 名，女性：2 名）に 5 回ずつ発話してもらい，実験データを収集した。実験では 1)非階層型手法，2)細分型手法，3)一括統合型手法，4)複合型手法の 4 種類を比較した。実験結果を図 8 に示す。

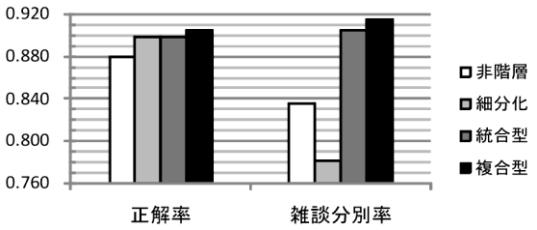

図 8：音声理解手法の精度

図 8 からわかるように，2 つの手法（細分型と一括統合型）を組み合わせた複合型手法が音声認識の正解率の面でも雑談検出率の面でも最高精度となり，提案手法の有効性が確認された。提案手法の一つである細分化型の雑談分別率が非階層型よりも劣るのは，編集距離による雑談判別の手法が，統合する認識器の数が増えた場合に適していないためであり，階層化手法の重要性も確認された。

(4)複数の認識器を利用した手形状認識とマルチモーダルインタフェース

非手形状を含む 7 つの形状パターンを用意し，学習データとして 4 名の被験者（男女 2 名ずつ）から，合計で 560 枚の画像を収集した。同様に 4 名の被験者（男女 2 名ずつ）について，560 枚の評価データを作成した。ここで，評価データの被験者のうち，男女 1 名ずつは学習データとは異なる人物である。さらに，逐次型学習器への追加学習用として，480 枚の画像を用意した。

正解率を表 1 に示す。表中の被験者 B と D が学習データには含まれていない人物である。ベースラインは，SVM のみ（複数の認識器を利用しない場合）の手法を意味する。実験結果から分かるように，提案手法では，特に未知の人物に対して精度が大幅に向上しており，その有効性が確かめられた。

表 1：手形状認識の精度

被験者	A	B	C	D
状態	既知	未知	既知	未知
ベースライン	0.85	0.77	0.88	0.73
提案手法	**0.95**	**0.90**	**0.91**	**0.87**

(5)言語と非言語情報を利用した対話の状態理解

10 名の学生から 4 名をランダムにピックアップし，5 分間の談話を収集した．収録にあたっては，事前に特定の話題を告知した。このような方法で 10 対話分収集し，2 名のアノテータにより，人手で盛り上がり度をタグ付けした。その結果，617 個の盛り上がり発話と 1099 個の非盛り上がり発話が得られ，このデータセットを基に 10 分割交差検定で評価した。

表 2 に実験結果を示す。表中の「統合」とは，言語情報と非言語情報の両方を組み合わせた手法（提案手法）を意味する。実験結果から分かるように，片方のみを利用した場合と比較して，提案手法は F 値の面で上回っており，提案手法の有効性が確認された。

表 2：盛り上がり推定の精度

手法	適合率	再現率	F 値
言語のみ	0.505	0.459	0.481
非言語のみ	0.478	**0.596**	0.531
統合	**0.535**	0.545	**0.540**

(6)画像を用いた人物識別

①属性情報を利用した人物識別

7 名の被験者に対して，40 日間の顔画像と衣服情報および時間情報を収集した。そのうち，訓練データとして 875 画像用意し，別の 175 画像を用いて評価した。隠れの状態として，サングラスによって目が隠れている場合とマスクによって鼻と口が隠れている場合の 2 種類を用意した。

実験結果を表 3 に示す。「併用」とはサングラスとマスクの両方を着用した場合を意味する。「顔のみ」とは衣服と時間特徴を利用しなかった場合である。「顔のみ」の手法では，「併用」状態では，顔のすべての特徴が隠されているため，人物識別は不可能である。顔の一部が隠れた場合であっても，衣服などの文脈特徴を利用することで精度が向上しており，提案手法の有効性が確認された。

表 3：文脈情報の有効性検証

	サングラス	マスク	併用	平均
顔のみ	0.697	0.880	−	−
提案手法	**0.966**	**0.926**	**0.977**	**0.956**

一方で，特徴量ごとにその有効性を検証すると，衣服特徴は効果的であったが，時間特徴については大きな精度向上には貢献しなかった。今後は他の特徴量の導入なども視野に入れる必要がある。

②頭上画像を利用した人物識別

実験の被験者は 8 名であり，そのうち 6 名が男性で，2 名が女性であった。被験者ごとに 30 枚の画像を撮影し，240 枚からなる実験データセットを構築した。この実験データを 10 分割交差検定で評価した。

実験結果を表 4 に示す。表中の「＋」は特徴量の組み合わせを意味している。最も精度が高くなったのは．つむじを除く特徴の組み合わせで，92.5%であった。

単独の特徴量としては体型が最も高い精度を算出した。一方で，体型特徴は，衣服などの変化に脆弱な特徴であることも実験によって確認している。実際に，ダウンジャケ

ットのような厚めの服を着てしまった場合では，その衣服情報に引きずられ，体型情報を大きく見積もってしまい，別の人物と誤識別する場合があった。一方で(6)-①で示したように，衣服情報は人物を識別するために有効な特徴であり，今後はその有効利用が必要になる。髪の色は最も脆弱な単体特徴であった。これは，日本人の多くが黒髪であることに起因する。つむじ特徴が有効に機能しなかった原因は，つむじ特徴自体の抽出に失敗する場合が多いことが挙げられる。つむじ特徴は他の特徴に比べて，本質的に頑健であると考えられるため，今後はより精度の高いつむじ特徴の抽出方法が必要となる。

表４：人物識別の精度

特徴	精度
体型	0.888
髪の色	0.283
髪型	0.633
つむじ	0.646
体型＋髪色	0.892
体型＋髪型	0.917
体型＋つむじ	0.888
髪型＋髪色	0.629
髪型＋つむじ	0.667
体型＋髪型＋髪色	**0.925**
体型＋髪型＋つむじ	0.883
体型＋髪色＋つむじ	0.888
髪型＋髪色＋つむじ	0.663
すべて	0.888

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Md. Khalilur Lhaman and Tsutomu Endo, Recurrent Neural Network Classifier for Three Layer Conceptual Network and Performance, JOURNAL OF COMPUTERS, 査読有, Vol.5, No.1, 2010, pp.40-48

〔学会発表〕（計１８件）

①横山貴彦，嶋田和孝，遠藤 勉．複数人談話における言語情報と非言語情報を利用した盛り上がり判定，言語処理学会第 18 回年次大会(NLP2012)，2012 年 3 月 14 日，広島市立大学.

② Kazutaka Shimada, Ryosuke Muto and Tsutomu Endo. A Combined Method Based on SVM and Online Learning with HOG for Hand Shape Recognition, The 2nd International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2011), SS4-3, 2011年11月21日，蘇州大学（中国).

③Ryota Nakatani, Daichi Kouno, Kazutaka Shimada and Tsutomu Endo. A Person Identification Method Using a Top-view Head Image from an Overhead Camera, The 2nd International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2011), SS4-1, 2011年11月21日，蘇州大学（中国).

④Kazuaki Komatsu, Kazutaka Shimada and Tsutomu Endo. A person identification method using facial, clothing and time feature, The 2nd International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2011), GS4-2, 2011年11月20日，蘇州大学（中国).

⑤Takahiko Yokoyama, Kazutaka Shimada and Tsutomu Endo. A Hierarchical Multiple Recognizer for Robust Speech Understanding, The Pacific Rim International Conference on Artificial Intelligence (PRICAI 2010), 2010年8月31日，ノボテルホテル大邱（韓国）.

⑥Kazutaka Shimada, Suguru Ishikawa and Tsutomu Endo. Web image retrieval for abstract queries using text and image information, The Fifth Asia Information Retrieval Symposium (AIRS 2009), 2009年10月22日，北海道大学

⑦Ryosuke Tadano, Kazutaka Shimada and Tsutomu Endo. Effective construction and expansion of a sentiment corpus using an existing corpus and evaluative criteria estimation, The 11th Conference of the Pacific Association for Computational Linguistics(PACLING2009), 2009年9月3日，北海道大学.

〔その他〕
ホームページ等
http://www.pluto.ai.kyutech.ac.jp/plt/endo-lab/index.html

６．研究組織
(1)研究代表者
遠藤　勉（ENDO TSUTOMU）
九州工業大学・大学院情報工学研究院・教授
研究者番号：１０１１２２９４
(2)研究分担者
嶋田　和孝（SHIMADA KAZUTAKA）
九州工業大学・大学院情報工学研究院・助教
研究者番号：５０３４６８６３