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２５年　５月３１日現在

機関番号：５２６０１
研究種目：基盤研究（Ｃ）
研究期間：2010～2012
課題番号：２２５００１７０
研究課題名（和文）　データマイニングを用いた日本語構文自己再編モデルによる手書き文章自動認識の研究
研究課題名（英文）　A Study on A Recognition Algorithm of Handwritten Japanese Document Using Syntax Self-Organization Model With Data Mining
研究代表者
鈴木　雅人（SUZUKI MASATO）
東京工業高等専門学校・情報工学科・教授
研究者番号：50290721

研究成果の概要（和文）：
一般の手書き文章では，字形に筆者の癖が強く現れ，また文章自体が日本語構文に合致しない場合が多いため，高い認識精度を実現するのが困難であった．本研究では，日本語構文の変遷・誤用に対する対策として，自然言語処理やデータマイニングを用いた，日本語構文解析の自己組織化モデルと，筆者の癖などに対応可能な標準パタン作成のための学習パタンの自己生成に関する研究を行った．その結果，従来の方法に比べて，手書き文章の認識精度を改善することができた．

研究成果の概要（英文）：
It is difficult to realize high recognition accuracy in Japanese handwritten document recognition, since a writer's peculiarity appeared strongly in type and the text itself don't agree in Japanese syntax. In this research, we propose the self-organization model of Japanese syntax using natural language processing a data mining, and the self-organization model of learning patterns of chinese character recognition. As a result, the recognition accuracy of Japanese handwritten document is improved compared with the conventional method.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,100,000	330,000	1,430,000
2011 年度	900,000	270,000	1,170,000
2012 年度	600,000	180,000	780,000
年度			
年度			
総　計	2,600,000	780,000	3,380,000

研究分野：　総合領域
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス
キーワード：パターン認識　機械学習　データマイニング　自然言語処理

１．研究開始当初の背景

日本語が危ないと言われて久しい．日本語学の専門家の中には，「文法的・構文的に誤った日本語が乱用されておりこのままでは日本語という言語が崩壊する」と指摘する学者がいる一方，「日本語が時代とともに変遷している」と考える専門家も多い．本研究の申請者らが専門とするパタン認識や機械学習の立場から，日本語学に関する議論をするのは

難しいが，日本語が日々変遷しており，それが社会で容認されているという現実がある以上，日本語を対象としたパタン認識技術も，日本語の変遷に追従して行く必要があることは言うまでもない．申請者らは長年，日本語の手書き文書の自動認識に関する研究を行っており，99.3%以上の認識精度を実現できるアルゴリズムも提案しているが，我々が通常書くような文字は，字形の多様性のため，十分な認識精度は得られていないのが現状である．また，手書き文章の認識精度を高めるため，記入枠指定，誤認識文字の手動訂正インターフェース構築などの研究も行われており，特に申請者らは，手書き宛名の自動認識に関する研究のほか，アンケートの自由記述欄の自動認識方法として，日本語文章の構文解析とWeb 検索を併用した手法も提案している．しかしいずれの場合においても，文章の内容を制限することによって認識精度を高める手法の提案であり，一般の手書き文章に対する高精度な自動認識アルゴリズムは確立されていないのが現状である．一般の手書き文章を高精度に自動認識するためには，個々の文字認識結果に対して構文解析・形態素解析を行うことが誤認識訂正に有効であるとされているが，この誤認識訂正法が有効にはたらくためには，個々の文字認識精度がある程度高いことと，認識対象の文章が，日本語の構文にかなった文章であることが大前提となる．しかし，前述のように日本語の文法が日々変遷していること，手書き文字の字形が筆者によって多様であることを考慮すると，これら２つの前提条件をクリアするためには，何らかの解決策を検討する必要がある．

２．研究の目的

前述のような問題を解決し，手書き文章認識の精度を改善するために，パタン認識技術に，自然言語処理・機械学習・データマイニングの研究成果を融合し，手書き文書の自動認識において，誤認識訂正に用いる日本語構文解析モデルに対して機械学習を適用し，解析モデルの漸次学習・自己組織化することによって，日本語の変遷に対応できる誤認識訂正モデルを確立する．また，個々の文字認識精度は，あらかじめ用意する学習パタンに依存するが，書き手が書いた数種類の文字と対応する標準パタンとの差異から書き手の癖などを抽出し，データマイニングを適用することで筆者の癖を吸収できるような学習パタンを全字種に対して自己生成し，筆者の癖を推定・考慮した文字認識方式が確立する．

これらの目的を達成するため，本研究では，(1) 誤認識訂正のための日本語構文自己組織化モデルの開発と，(2)データマイニングを活用した学習パタンの漸次生成アルゴリズムの開発とに焦点を絞って研究を行う．

一般の手書き文章の自動認識では，個々の文字認識結果に対して構文解析・形態素解析を行い，単語の切り分けの妥当性・品詞・係り受け関係などを調査して，誤認識箇所の特定・訂正を行っている．しかし文法的に正しくない文章では，このような誤認識訂正法に限界があるため，それぞれの単語の品詞や係り受け関係が，社会で容認されうるものかどうかを検証し，その結果をフィードバックすることで，誤認識訂正のための日本語構文自己組織化モデルを開発する．

また，データマイニングを活用した学習パタンの漸次生成アルゴリズムの開発に関しては，申請者らは既に，多様な手書き文字の認識に対応できる学習パタンの自動生成法について成果をまとめている．しかし，手書き文字の多様性を容認すると学習パタンのばらつきは連動して大きくなるため，統計的な識別手法では大きな問題となる．本研究では，認識対象が変わるたびに，それらに適した学習パタンを自動生成するのではなく，文字パタンを認識して行く過程において，書き手の癖を検知し，学習パタンに対してデータマイニングを適用する手法を検討し，学習パタンの漸次自動生成アルゴリズムの開発を行う．

３．研究の方法

研究の最終的な目的は，自由形式の手書き文章の高精度な自動認識アルゴリズムを考案し，システムとして実装することである．この研究目的を達成するためには，研究目的に示したように，(1)誤認識訂正のための日本語構文自己組織化モデルの開発，(2)データマイニングを活用した学習パタンの漸次生成アルゴリズムの開発，の２つの課題に取り組む必要がある．

誤認識訂正のための日本語構文自己組織化モデルの開発については，まず，既存の研究成果をもちいて本研究の目的の実現可能性について検証し，問題点の洗い出しを行い，その結果をもとに，日本語構文自己組織化モデルの構築方法について検討する．特に構文解析などで検出される誤った表現に対して，そのような表現が現社会で許容されるかどうかの判断基準が必要となるため，その基準についても検討する．判断基準の１つの案として，Ｗｅｂ上に散在する文章の中から特定の文章を収集し，それらの情報を手がかりに判断する方法を検討する．検討したアルゴリズムを手書き文章認識システムに組み込んで認識実験を行い，日本語構文モデルの変化を検証し，本申請の当初の目的を達成しうる動作ができているかどうかを検証する．

データマイニングを活用した学習パタンの漸次生成アルゴリズムの開発では，従来のよ

うに多様な文字パタンに対応できる学習パタンを生成するのではなく，認識対象文字の癖・特性に応じて学習パタンが動的に適応できるような，学習パタンの漸次生成法を検討する．

そのための第一段階として認識対象文字の癖・特性について議論し，それらを数値化する方法を検討する．一般に漢字を書くためには，おおまかに８つの技法が必要とされており，それらは「永字八法」として知られている．本研究ではこれらの技法に注目し，標準パタンと認識対象パタンとの差から認識対象文字の癖・特性を抽出することを検討する予定である．また，幾つかの認識対象文字から抽出した癖・特性がわかれば，データマイニングの技術を活用することにより，他の字種についても癖・特性を予測し，より高精度に認識できると考えている．本研究では，このような学習パタンの動的構成法を検討し，手書き文書認識システムに組み込む．ただし，学習パタンの自動生成によって，認識処理時間が大幅に長くなることが容易に予測できるため，実際に実装を行った段階で，処理速度をどの程度改善すべきかを見積もり，サーバスペック・学習パタン生成アルゴリズムの改善など，あらゆる面から，実用に耐えるシステムの実現を目指す．

４．研究成果

本研究では，誤認識訂正のための日本語構文の自己組織化モデルの開発と，データマイニングを活用した文字認識用学習パタンの漸次生成アルゴリズムの開発を目的として研究を行った．

日本語構文の自己組織化モデルの検討については，まず，既存の構文解析システムや形態素解析システムを用いた誤認識訂正の可能性について模索した．その結果から，文法的に正しくない場合も既存システムで解析ができるため，解析結果から誤認識箇所を特定し訂正するのは困難であることがわかった．次に，その結果を受けて，日本語構文の自己組織化モデルの検討を行った．認識処理後の構文解析結果から誤認識箇所を特定し訂正する方法について検討し，学生のレポートなどに代表されるように，筆者の表現上の癖を構文解析結果から抽出することにより誤認識箇所の検出および訂正するための枠組みを完成させることが出来た．提案アルゴリズムを手書き文書認識システムに組み込み評価実験を行ったところ，誤認識箇所検出・訂正に有効であることがわかり，この枠組みを用いて，日本語構文の自己組織化モデルを作り上げることができた．

文字認識用学習パタンの自動生成については，これまで提案されてきた手法においては，手書き文字に対応可能な多様なパタンの生成は可能でも，その中から認識精度改善に有効なパタンを選別することが難しく，そのため標準パタンの分布がぼけてしまい，認識精度改善には至っていなかった．本研究では，永字八法に着目し，手書き故に癖が出やすい，「横画」や「右はらい」などに着目し，筆者の癖を抽出することで，筆者毎に有効な学習パタンの自動生成アルゴリズムを検討した．しかし，検討した手法では，膨大な変形パタンを生成し，時間のかかる選別方法を経て，その中から有用なパタンを絞り込むため，パタン生成時に筆者の癖情報を適用する範囲に制限を設けることで，パタン選別をある程度高速化することができた．尚，この提案アルゴリズムにより当初の研究目的は達成できたと考えているが，実際に生成された文字画像の特徴量を調査すると，筆者の癖に合わせてパタン生成するよりも，むしろ，文字画像の変形をある種の特徴ととらえ，その特徴の分布に適切な密度関数を当てはめて識別を行う方が，処理時間的な問題点をクリアできるのではないかとの考えに至った．この考え方により，当初の予定で懸念材料となっていた，学習パタン生成にかかる処理時間が指数関数的に増えることに対しても，大幅な改善を施すことができた．

以上，２つの課題について取り組み，アルゴリズムを実装し，実験データを用いて検証を行った．その結果，従来の手書き文書認識に比べて処理内容が増えているために認識時間は長くなるが，文章認識精度を大幅に改善できることが検証できた．また，これらの研究成果を研究会などで公表した．

５．主な発表論文等

〔雑誌論文〕（計２件）

1. 鈴木雅人，北越大輔，“低速通信回線利用を前提とした文字認識における文字画像処理方式の一検討”，東京高専研究報告書　　第 42(1)号，査読有，pp.103-106，2010.
2. Uthai Phommasak, Daisuke Kitakoshi, and Hiroyuki Shioya : An Adaptation System in Unknown Environments Using a Mixture Probability Model and Clustering Distributions, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 16, No. 6, pp. 733-740, 2012.

〔学会発表〕（計６３件）

1. Daisuke Kitajima, Daisuke Kitakoshi, and Masato Suzuki, “ENGLISH VOCABULARY LEARNING SUPPORT SYSTEM

BASED ON STOCHASTIC WORD PROFICIENCY MODEL ESTIMATING ELUSIVE WORDS FOR STUDENTS”, Proceedings of the Second International Conference on Digital Enterprise and Information Systems (DEIS2013), 2013 年 3 月 4 日, Kuala Lumpur (Malaysia).

2. 松本章代, 今村真浩, 小西達裕, 高木朗, 小山照夫, 三宅 芳雄, 伊東幸宏, “日本語ウェブ文書を対象とした 10 年間の実態調査 (2002 年～2012 年)”, 第 5 回データ工学と情報マネジメントに関するフォーラム, 2013 年 3 月 3 日, 福島県郡山市ホテル華の湯.
3. 鈴木雅人, 北越大輔, 松本章代, ″JohnsonSU 分布を用いた手書き文字認識用識別関数の改良″, 電子情報通信学会 2013 年総合大会, 2013 年 3 月 20 日, 岐阜大学.
4. 鈴木雅人, 北越大輔, 松本章代, ″JohnsonSU 分布を用いた異字種パタン検出に関する一検討″, 情報処理学会第 75 回全国大会, 2013 年 3 月 8 日, 東北大学.
5. 浅井 彬弘, 松本 章代, “タブレット端末向け添削アプリケーションの開発”, 平成 24 年度情報処理学会東北支部研究会, 2013 年 2 月 13 日, 東北学院大学.
6. 鈴木雅人, 北越大輔, 松本章代, “歪度最大基準に基づく特徴選択法による低品質手書き文字認識手法の検討”, 電子情報通信学会パターン認識・メディア理解研究会, 2013 年 1 月 23 日, 京都大学.
7. 市川詠一, 鈴木雅人, 北越大輔, 西村亮, “採点ミス誤り自動検出による入試採点支援システムの開発”, 第 4 回大学コンソーシアム八王子学生発表会, 2012 年 12 月 8 日, 八王子市学園都市センター.
8. 北島大資, 北越大輔, 鈴木雅人, “苦手単語を推測する単語特徴量?習得モデルを用いた英単語学習法”, 生命ソフトウェアシンポジウム 2012, 2012 年 11 月 23 日, 室蘭工業大学.
9. 岡野卓矢, 北越大輔, 鈴木雅人, “介護予防運動における強化学習ロボットの活用に関する検討”, 生命ソフトウェアシンポジウム 2012, 2012 年 11 月 23 日, 室蘭工業大学.
10. 櫻井 優, 坂本泰伸, 松澤 茂, 武田敦志, 松本章代, 富樫 敦, 柏葉俊輔, “高齢者の QOL 向上を目指した Android システムの実証実験の結果報告”, 情報処理学会第 152 回 DPS・第 85 回 GN・第 57 回 EIP 合同研究発表会, 2012 年 9 月 13 日, 尾道公会堂別館.
11. 松本章代, 高橋光一, “科学技術文書の論理性を推敲できる文章作成教育システムの構築”, 教育システム情報学会第 37 回全国大会, 2012 年 8 月 22 日, 千葉工業大学.
12. 鈴木雅人, 北越大輔, “部分的正規分布に基づくパタン類別法による入試採点誤り検出の検討”, 電子情報通信学会データ工学研究会, 2012 年 8 月 1 日, 名古屋大学.
13. 鈴木雅人, 松石浩輔, 北越大輔, 松本章代, “確率ネットワークを用いた手書き文書認識の後処理方式の検討”, 信学 2012 年総合大会, 2012 年 3 月 22 日, 岡山大学.
14. 松石浩輔, 鈴木雅人, 松本章代, 北越大輔, “文章表現の癖抽出に基づく手書き文章認識の後処理方式の検討”, 信学 2012 年総合大会, 2012 年 3 月 22 日, 岡山大学.
15. 向山和宏, 鈴木雅人, 北越大輔, “入試採点支援における採点責任者業務支援システムの開発”, 信学 2012 年総合大会, 2012 年 3 月 24 日, 岡山大学.
16. 榎本大義, 北越大輔, 鈴木雅人, “一般道における渋滞緩和・解消を図る交通信号機制御システムに関する研究”, 情報処理学会第 74 回全国大会, 2012 年 3 月 6 日, 名古屋工業大学.
17. 岡野卓矢, 北越大輔, 鈴木雅人, “Human-Agent Interaction を導入した強化学習エージェントによる人工知能デモシステム”, 情報処理学会第 74 回全国大会, 2012 年 3 月 6 日, 名古屋工業大学.
18. 相良光志, 北越大輔, 鈴木雅人, “ベイジアンネットによる Web ブックマーク選択モデルを用いたブックマーク推薦法”, 情報処理学会第 74 回全国大会, 2012 年 3 月 6 日, 名古屋工業大学.
19. 山崎大地, 北越大輔, 鈴木雅人, “相互作用型階層強化学習システムによるエージェント集団の共存期間伸長に関する検討”, 情報処理学会第 74 回全国大会, 2012 年 3 月 6 日, 名古屋工業大学.
20. 北島大資, 北越大輔, 鈴木雅人, “語彙学習用確率モデルを利用した英単語学習支援システムに関する一考察”, 第 64 回人工知能学会先進的学習科学と工学研究会, 2012 年 3 月 13 日, プラザ淡路島.
21. 田中功太, 北越大輔, 鈴木雅人, “教員の指導法改善を目的した授業評価・習熟度関連性モデルによる知識発見支援システム”, 第 64 回人工知能学会先進的学習科学と工学研究会, 2012 年 3 月 13 日, プラザ淡路島.
22. 和歌崎修平, 北越大輔, 鈴木雅人, “精

度保証と補正を行うベイジアンネット上の近似確率推論法に関する研究”，電子情報通信学会 NC 研究会，2012 年 1 月 26 日，はこだて未来大学.

23. 安藤大輝，北越大輔，鈴木雅人，“強化学習エージェントの方策情報ベクトル表現を用いた学習効率化に関する研究，電子情報通信学会 NC 研究会，2012 年 1 月 26 日，はこだて未来大学.
24. 伊賀篤史，鈴木雅人，北越大輔，“拡大法を用いた円形道路標識認識”，第 3 回大学コンソーシアム八王子学生発表会，2011 年 12 月 3 日，八王子市学園都市センター.
25. 海老原昌吾，西村亮，鈴木雅人，北越大輔，“緊急自動車の接近を検知する聴覚障害者支援アプリケーションの開発”，第3回大学コンソーシアム八王子学生発表会，2011 年 12 月 3 日，八王子市学園都市センター.
26. 笹岡耕地，鈴木雅人，北越大輔，“入試採点支援システムにおける数式の正誤判定”，第 3 回大学コンソーシアム八王子学生発表会，2011 年 12 月 3 日，八王子市学園都市センター.
27. Daisuke Kitakoshi, Shuhei Wakasaki, and Masato Suzuki, “A Probabilistic Reasoning Algorithm for Bayesian Networks by Simplifying Their Structures”, Proceddings of the 2011 IEEE International Conference on Granular Computing, 2011年11月8日, Taiwan.
28. Naoto Osaka, Daisuke Kitakoshi, and Masato Suzuki, “A Reinforcement Learning Method to Improve the Sweeping Efficiency for an Agent”, Proceddings of the 2011 IEEE International Conference on Granular Computing, 2011年11月8日, Taiwan.
29. Daisuke Kitakoshi, Ryunosuke Miyauchi, and Masato Suzuki, “A Study on Reinforcement Learning System for Agents to Acquire Cooperative Behavior in Gap-Widening Situations”, Proceeding of 2011 IEEE Workshop on Robotic Intelligence in Informationally Structured Space, 2011年4月11日, France Paris.
30. 鈴木雅人，大久保貴博，北越大輔，松本章代，“永字八法に基づく手書き文字認識用辞書の動的構成法”，2011 年電子情報通信学会総合大会,2011年3月16日，東京都市大学.
31. 和歌崎修平，北越大輔，鈴木雅人，“ネットワーク構造簡略化を用いた確率推論手法による意思決定支援”，第 38 回知能システムシンポジウム，2011 年 3 月 16 日，神戸大学.
32. 西山遥，北越大輔，鈴木雅人，“ベイジアンネットの段階的構造学習法における適切なパラメータ設定に関する研究”，第38回知能システムシンポジウム,2011年3月16日，神戸大学.
33. 大澤翔吾，萩原奈央，北越大輔，鈴木雅人，“入試採点システムにおける手書き文字の特徴量の類似性を用いた採点ミス検出アルゴリズム”，第 38 回知能システムシンポジウム，2011 年 3 月 16 日，神戸大学.
34. 大坂直人，北越大輔，鈴木雅人，“仮想ロボットによる掃引作業計画の効率化を目的とした強化学習法”，第 38 回知能システムシンポジウム，2011 年 3 月 16 日，神戸大学.
35. 丸田拓和，西村亮，北越大輔，鈴木雅人，“エージェントの行動獲得過程を効果的に支持するデモシステムの開発とその評価”，第 61 回人工知能学会先進的学習科学と工学研究会,2011年3月13日，山口県長門市湯元観光ホテル西京.
36. 鏡沼悠太，北越大輔，鈴木雅人，“授業計画・習熟度関連性モデルを用いた教員の知識発見に基づく対話型指導法改善支援システム”，第 61 回人工知能学会先進的学習科学と工学研究会，2011 年 3 月 13 日，山口県長門市湯元観光ホテル西京.
37. 松本章代，Martin Duers，“可読性の指摘を行うプログラミング教育システムの開発－反復構造の自動検出による関数化の指摘”，情報処理学会第 73 回全国大会，2011 年 3 月 3 日，東京工業大学.
38. 本間皇成，松本翔太，松本章代，松原俊一，Martin J. Duerst，“チームプログラミングを可能とした教育支援システムの開発”，情報処理学会第 73 回全国大会，2011 年 3 月 3 日，東京工業大学.
39. 北越大輔，和歌崎修平，鈴木雅人，“Conditioning と Loopy-BP を用いた確率的意思決定手法”，電子情報通信学会 NC 研究会，2011 年 1 月 24 日，北海道大学.
40. 山崎泰幸，北越大輔，鈴木雅人，“手書き数式認識を用いた基礎数学学習支援システムの開発”，第 2 回大学コンソーシアム八王子学生発表会，2010 年 12 月 4 日，八王子市学園都市センター.
41. 萩原奈央，北越大輔，鈴木雅人，“入試採点支援システムの開発”，第 2 回大学コンソーシアム八王子学生発表会，2010 年 12 月 4 日，八王子市学園都市センター.
42. 山崎大地，北越大輔，鈴木雅人，“相互

作用型階層強化学習システムのマルチエージェント環境における特性評価”, 第 20 回インテリジェントシステムシンポジウム, 2010 年 9 月 25 日, 首都大学東京.

６．研究組織

(1)研究代表者

鈴木　雅人 (SUZUKI MASATO)
東京工業高等専門学校・情報工学科・教授
研究者番号：50290721

(2)研究分担者

(3)連携研究者

北越　大輔 (KITAKOSHI DAISUKE)
東京工業高等専門学校・情報工学科・
准教授
研究者番号：50378238

松本　章代 (MATSUMOTO AKIYO)
東北学院大学・教養学部・講師
研究者番号：40413752