様式 C-19

科学研究費補助金研究成果報告書

平成　23 年　6 月　1 日現在

研究種目：若手研究（A）
研究期間：　2007～2009
課題番号：19680007
研究課題名（和文）　主辞駆動句構造文法のための統計同期文法による機械翻訳

研究課題名（英文）　Machine Translation with Probabilistic Synchronous Head-driven Phrase Structure Grammars

研究代表者
二宮 崇（NINOMIYA TAKASHI）
愛媛大学・大学院理工学研究科・准教授
研究者番号：20444094

研究成果の概要（和文）：本研究では言語学的文法に基づく機械翻訳を実現するために、主辞駆動句構造文法(HPSG)のための同期文法(二言語のための対文法)のモデル化、同期 HPSG 文法の作成、HPSG 構文木付きの英日テキストデータの作成、および同期 HPSG 文法による機械翻訳の実験を行った。また、機械翻訳のための基礎研究として、HPSG のための決定性構文解析の研究、学習時間が短くかつ高精度なオンライン学習理論の研究、自動的に難しい英単語を予測し、その訳を表示する英文読解支援の研究を行った。

研究成果の概要（英文）：We studied a model of synchronous Head-driven Phrase Structure Grammars for machine translation. We developed an experimental synchronous HPSG grammar, and parallel texts annotated with HPSG parse trees. The performance of machine translation with synchronous HPSG was evaluated on newspapers as preliminary experiments. As fundamental studies for machine translation, we also studied deterministic parsing for HPSG, fast and accurate online learning, and reading support of second languages.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	3,000,000	900,000	3,900,000
2008 年度	2,800,000	840,000	3,640,000
2009 年度	2,800,000	840,000	3,640,000
年度			
年度			
総　計	8,600,000	2,580,000	11,180,000

研究分野：自然言語処理
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、機械翻訳、多言語処理、HPSG、構文解析、オンライン学習、英文読解支援

１．研究開始当初の背景

近年、計算機の性能が大きく向上し、また、機械翻訳のための言語的資源の質的量的な増大、および高性能な機械学習器の利用が可能になったことにより、より洗練された機械翻訳手法を用いることが可能となった。機械翻訳の研究は過去数十年にわたって行われてきたが、90 年代半ば以前は、パラメータを

人手により調整するルールによる手法、構文構造を変換する手法、用例に基づく手法および言語学的な文法を用いた意味表現を介する機械翻訳手法を中心に研究がなされてきた。しかし、これらの手法では、ルール作成やパラメータ調整を人手により行うため、ルール間の相互作用による副作用を考慮しながらルールを追加・削除し、パラメータを調整することが難しく、その結果、大規模なシステムを構築することや、性能を向上するための改良が非常に困難であった。言語学的な文法を用いた意味表現を介する機械翻訳はドイツの DFKI の VERB MOBIL プロジェクトで過去行われてきたが、実用的なレベルにまでは達していない。これは、実テキストを広く解析できる大規模文法の開発が困難であったこと、および過度に一般化された意味表現を介して言語間のマッピング規則を構築することが困難であったためと考えられる。90 年代にはいって、統計機械翻訳と呼ばれる統計モデルに基づく機械翻訳が IBM の Brown らにより提案され、機械翻訳の業界で注目されている。これは、音声認識や形態素解析でよく用いられている統計的モデルを機械翻訳に応用した機械翻訳方式であり、教師あり・なし機械学習によるルール・パラメータの自動学習が行えるため人手の介在が少なくすみ、上述の副作用を伴わない特長をもつ。しかしこの手法は、単語の翻訳・移動・マッピングを単位とした翻訳手法であり、言語学的制約や句構造を単位とした機械翻訳手法ではないため、日英翻訳のように構文構造が大きく異なる言語間では高精度化が非常に難しい。

ここ数年、統計機械翻訳に句構造をとりこんだ研究や、逆にルールベースや用例ベースに統計モデルを導入する手法が研究されており、この一般化として、統計的同期文法が注目を浴び、研究され始めている。同期文法は、二つの言語を記述する句構造文法間にマッピングを与えた文法であり、数学的によく定義されたモデルになっているため統計モデルと相性が良く、文法的な制約により自然な翻訳がなされることが期待されている。同期文法は 1960 年代 Aho と Ullman により提案されており古くから存在するが、この数年で注目を浴びている理由としては、高速で精度の高い構文解析器が一般に利用可能になったこと、文単位で対応がつけられた 2 言語間の翻訳テキストが利用可能になったこと、および自然言語処理で統計モデルの研究が大きく発展したことが大きな理由と考えられる。しかしながら、LTAG と呼ばれる文法のための同期文法を除く既存の同期文法は、CFG のための同期文法がほとんどであり、言語学的に厳密に定義された文法のための同期文法はまだ提案されていない。言語学的に厳密に定義された文法を用いることで、より文法的な句構造間の対応付けや、より文法的な文の生成が可能になることが期待される。

２．研究の目的

本研究は、言語学的に定義された主辞駆動句構造文法(HPSG)のための同期文法をモデル化し、実際に開発することにより、より洗練された機械翻訳を実現することを目的とする。HPSG は言語学的に厳密に定義された語彙化文法であり、そのため、HPSG のための同期文法は、CFG のための同期文法よりも、より文法的な句構造間の対応付けや、より文法的な文の生成が可能となることが期待される。本研究では、同期 HPSG による機械翻訳の実現のため、機械翻訳に関する基礎理論の研究、言語資源とツールの構築、同期 HPSG の理論化、作成、および機械翻訳による性能評価を行う。

まず、基礎理論について次の(1)～(3)の研究を行う。

(1) オンライン学習

機械翻訳における学習を容易とするためには、より高精度でかつ高速な学習を行える機械学習の理論が必要となる。本研究では、オンライン学習の高精度化を目的とする。オンライン学習は、データが逐次的に入力され、逐次的に最適化を行う学習手法の総称であり、データ全体に対して最適化を行う学習はバッチ学習と呼ばれる。オンライン学習はバッチ学習に比べ精度の面で劣るものの学習速度とメモリ効率の点で非常に優れていると考えられている。本研究では、バッチ学習と同等以上の精度を達成するオンライン学習手法の実現を目的とする。

(2) 決定性 HPSG 構文解析

機械翻訳に要求される技術は構文解析に必要となる技術と類似しており、高速かつ高精度な構文解析技術は機械翻訳においても有用である。特に、パラレルコーパスを同期文法で構文解析を行うことにより、同期文法による単語対応付け(単語アライメント)が実現でき、従来の単語と句単位のアライメント技術よりも精度の高いアライメントが実現できることが期待される。本研究では、高速な HPSG 構文解析を実現する決定性構文解析の研究を行う。

(3) 英文読解支援

機械翻訳に関連する技術として、自動的に難しい単語を検出し、その訳語を表示する英文読解支援の研究を行う。

機械翻訳を行うため、次の(4)と(5)の研究・開発を行う。

(4) 対訳コーパスと素性構造処理ツールの開

発
同期 HPSG による機械翻訳を実現するためには、同期 HPSG の正解構文木集合が存在することが望ましい。本研究では、その実現のために、Penn Treebank と呼ばれる構文木付きコーパスに含まれる英字新聞 Wall Street Journal に対する対訳を作成することを目的とする。また、HPSG は、素性構造と呼ばれる単一化が定義されたグラフ構造を用いて定義されているため、それらを計算機上で扱うためのツールを開発する。

(5) 同期 HPSG のモデル化、文法開発、評価
本課題の目標である同期 HPSG の理論化、文法開発、また、BLEU 等のスコアによる評価を行う。

３．研究の方法
本研究では、２．であげた(1)～(5)の目的に対し、以下の研究を行った。
(1) 多クラス識別問題におけるオンライン学習のための厳密な PA アルゴリズム
本研究では PA 戦略と呼ばれるオンライン学習の枠組における厳密解法による新しいオンライン学習手法を提案する。Crammer らによって提案された PA アルゴリズムは代表的なオンライン学習アルゴリズムであるが、多クラス識別問題においては、厳密な PA アルゴリズムの枠組みから外れた近似解法が用いられてきた。本研究は本来の PA アルゴリズムの厳密解をサポートクラスという概念を用いて導出し、これらに基づく識別関数の更新を効率的に行う為のアルゴリズムを提案する。サポートクラスを適切に定めることによって、更新後の分類器は、訓練データを正しく分類できるようになる。このアルゴリズムは、受け取った訓練データを一定のマージンで正しく分類することを強制するような損失関数を設計し、PA アルゴリズムの枠組みで用いることによって自然に得られる。本研究ではこの戦略における厳密解を閉じた式で導出することにより、新しいオンライン学習アルゴリズムを導出する。

(2) 単一化文法のための決定性構文解析
決定性構文解析は CFG 構文解析や依存構造解析においてよく研究されており、探索を決定的に行うため、高速に解析を行うことができ、また、大域素性を用いることができるため、完全な探索に近い精度の解析が実現出来る。しかし、HPSG などの単一化文法は制約により記述されるため、CFG のための決定性構文解析手法を単純に適用すると、制約違反により解析に失敗し、決定的に解析を進めることができない。本研究は、デフォルト単一化と呼ばれるほぼ失敗をすることがない特殊な単一化を用いることにより、単一化文法における決定性構文解析を実現する。

(3) 英文読解支援
本研究は、ユーザの英語レベルに合わせて英文に対し語義のアノテーションを自動的に与える手法を提案、および、ウェブ上で動作するシステムを開発する。システムは、登録ユーザのレベルに応じて、そのユーザにとって未知であると推測される英単語に日本語訳を自動的に付与する。システムが行うユーザの既知/未知単語予測には項目反応理論 (Item Reponse Theory, IRT) の一つであるラッシュモデルを拡張した確率モデルを用いた。ラッシュモデルにおいて定義される素性に加え、次の素性を確率モデルに追加した; ウェブコーパスの単語頻度、SVL12000 の単語難易度表。また、確率モデルの学習に数百人規模のユーザに対する語彙情報も用いた。

(4) 対訳コーパスとツールの開発
同期 HPSG のための機械翻訳を実現するためには、学習および評価のために、同期 HPSG による正解構文木集合が存在することが望ましい。しかし、そのような同期文法に対する正解データは存在しないため、直接開発するか、自動的な解析により得るしかない。しかし、自動的な解析により得られる構文木集合や単語アライメントは、高い頻度でエラーが含まれるため、良い学習がなされない可能性が高い。一方、Penn Treebank と呼ばれる構文木付きコーパスには、英字新聞 Wall Street Journal の記事に対する CFG 構文木が収録されており、これを用いて、HPSG 構文木の正解データがすでに開発されている。本研究では、この正解データに含まれる英文に対し、人手による対訳を行い、Penn Treebank の英日対訳テキストを開発する。すでに存在するパラレルコーパス(文対応付き対訳テキスト)に対し正解 HPSG 構文木を開発するという考え方もあり得るが、対訳テキストを作成するコストは正解 HPSG 構文木を開発するコストよりも低いため、すでに存在する正解 HPSG 構文木に対訳テキストを開発する戦略をとる。また、HPSG は、素性構造と呼ばれる単一化が定義されたグラフ構造を用いて定義されているため、素性構造を計算機上で扱うためのツールが必要となる。LiLFeS と呼ばれるプログラミング言語ツールにおいて素性構造を扱えるが、より容易に作成できるツールを開発することを行う。

(5) HPSG のための同期文法のモデル化、開発、評価
HPSG は LTAG と同じ語彙化文法であり、LTAG のための同期文法と同様に、語彙項目の対応付け、文法規則の対応付けによりモデ

ル化される。図 1 の下側は同期 HPSG による構文解析の結果を示している。同期文法の構文木は二言語の構文木の対で表現され、図で表されるように、単語を含めて各句構造は点線で示される対応付けがされている。同期 HPSG の文法規則は、二言語の文法規則対として定義され、文法規則の対応と、二言語間の句や語の線形順序を定義する。HPSG は極端に文法規則が少ない語彙化文法であり、十程度の数の文法規則しかもたないため、文法規則の対を記述するのは比較的容易である。対となる構文木は、対となる文法規則を対となる句構造に繰り返し適用することにより得られる。

HPSG の同期文法を得るために、日本語、英語の 2 言語の文法が必要となるが、現在、現実世界のテキストを十分に解析できる英語 HPSG 文法は存在するが、日本語 HPSG 文法は存在しない。これは計算機で解析可能な日本語の述語項構造の定義が難しく、既存の木構造付テキストを利用することが難しいことが原因であると考えられる。本研究では、英語文法と独立に日本語文法を作成するのではなく、英語 HPSG を獲得した手法であるコーパス指向文法開発を用いて、半自動的に獲得することを行う。図 1 はその獲得過程を示している。

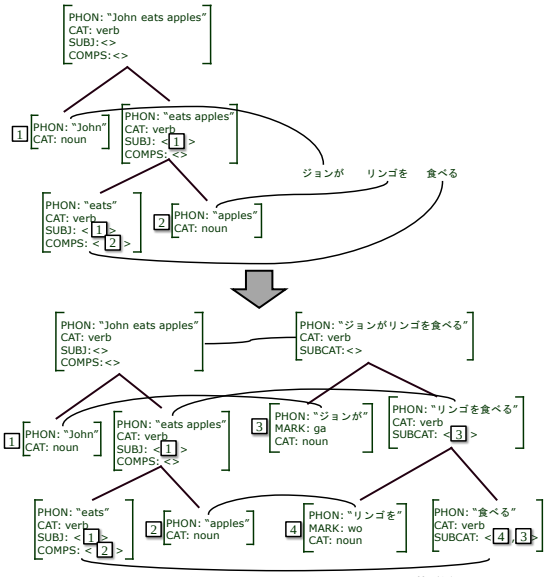

図 1: 同期 HPSG 構文木の獲得

まず、(4)で開発する文単位で対応がつけられた日英対訳テキストに自動的に単語単位の対応をつける手法（単語アライメント）を適用する(図 1 上側)。単語アライメントには GIZA++を用いる。(4)で開発する対訳テキストの英語側には英語 HPSG 文法による正解構文木が付与されているため、対応する日本語の文法規則を適用することにより、図 1 の下側のように半自動的に英日間の同期 HPSG の構文木が得られる。最後に、同期 HPSG 文法を用いて、機械翻訳における性能評価を行う。ただし、本研究においては、同期文法による構文木の構造的対応の性能について評価を行い、単語に対する対訳辞書や単語の翻訳に関する評価は将来の課題とする。

４．研究成果

(1) 多クラス PA アルゴリズムの効率的な厳密解法によるオンライン学習の研究成果

本研究で提案する多クラス PA アルゴリズムの効率的な厳密解法の実験を行った。20Newsgroups、Reuters、USPS のデータセットを用いて実験を行った。20Newsgroups、Reuters は文書分類、USPS は文字認識のタスクになっている。提案手法と、PA、パーセプトロン(Perc)、バッチ学習法である多クラスロジスティック回帰（LR）及び多クラス SVM (SVM) と比較した。誤識別率の結果を表 1 に所要時間を表 2 に示す。

表 1: 誤識別率(%)

	PA	提案手法	Perc	LR	SVM
USPS	6.82	**4.95**	6.45	4.98	4.83
Reuters	3.81	**2.96**	3.99	3.50	3.18
News20	21.02	**15.83**	22.63	17.79	15.94

表 2: 所要計算時間(秒)

	PA	提案手法	Perc	LR	SVM
USPS	5.5	**5.9**	1.5	53.6	68.5
Reuters	6.2	**6.4**	1.0	92.5	34.3
News20	15.8	**17.2**	7.9	405.6	75.7

実験結果より、提案手法は PA アルゴリズムと比べほぼ同じ速度でより高い精度を達成している。従来の主な学習手法であった SVM やロジスティック回帰のバッチ学習に比べ、非常に高速に学習し、精度もほぼ同等かそれ以上の性能を示した。結果として得られた更新を用いたアルゴリズムは優良であり、今後計算コスト的にも精度面においても優良な学習のためにオンライン学習が用いられうることが示唆できたと考えられる。この研究成果は、主要な国際会議である SIAM International Conference on Data Mining (SDM10)に採択された。

(2) 単一化文法のための決定性構文解析の研究成果

CFG における決定的構文解析が効率、精度の面から大きく注目されているが、決定的構文解析を単一化文法に対し適用することは困難であった。本研究では文法規則の適用にほとんどの場合において失敗しないデフォル

ト単一化を用いることでこの問題を解決した。表 3 は実験結果を表す。

表 3: 決定性 HPSG 構文解析の実験結果

	普通の単一化		デフォルト単一化	
	精度	時間(ミリ秒/文)	精度	時間(ミリ秒/文)
従来手法	86.9%	604		
提案手法 1	79.1%	122	**87.6%**	**256**
提案手法 2	87.0%	510	**88.5%**	**457**

実験により、決定的構文解析(提案手法 1)の精度はデフォルト単一化により 79.1%から 87.6%まで向上したことを確認した。また、その他に構文解析の失敗から回復する手法(提案手法 2)も実験を行い、88.5%の精度を達成することを確認した。この成果は、主要な国際会議である Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)に採択された。

(3) 英文読解支援の研究成果
本研究では英文の読解を支援するために、英文に対し自動的に語義のアノテーションを与えるシステムを提案した。学生 10 人に、12,000 語について単語の既知/未知のラベルを与えたデータを作成し、このデータを用いてシステムの性能を評価した。外部リソースとして、Google n-gram コーパスによる単語頻度、SVL12000 単語の単語難易度表、Smart.fm と呼ばれるコンピュータ支援語彙習得システムから得られる 675 人分の既知/未知の欠損データを用いる手法を提案し、評価した。実験により、およそ 100 単語に対する既知/未知の情報を得ることでおよそ 80%の精度で残りの単語の既知/未知を予測することが出来た。また、このようなシステムにおいてオンライン学習を用いることの有用性も確認できた。この研究成果は主要な国際会議である International Conference on Intelligent User Interfaces (IUI2010)に採択された。学習および評価のための正解データは本科研費の予算を用いて作成した。

(4) Penn Treebank の英字新聞 Wall Street Journal の英日対訳コーパスの開発成果
Penn Treebank の Wall Street Journal は全部で 49,208 文からなる。この一部となる 18,635 文については、NICT によってその対訳がすでに開発されている。本研究は残った 30,573 文のうち 11,482 文の対訳を作成した。この作業によって Penn Treebank の Wall Street Journal に対し 30,117 文の英日対訳対が利用可能となった。
作成した 30,117 文の英日対訳対に対し、統計機械翻訳のツールとしても有名な Moses を用いて、翻訳精度を測定した。25,946 文を学習用コーパス、2182 文をパラメータ調整用コーパス、1989 文をテストコーパスとして実験を行った。日本語の単語分割は MeCab を用いた。表 4 は仏英 WMT08 News Commentary (wmt08, 約 55,000 文)による Moses の性能と、本研究で開発した Wall Street Journal (wsj) に対する性能を評価した実験の結果を示す。

表 4: Moses による性能評価

	NIST	BLEU
wmt08	7.1037	0.2523
wsj	4.2729	0.1193

wsj はスコアが低い傾向にあるが、仏英よりも英日の方が難しい、学習コーパスが小さい、日本語の単語分割が適切でない、wsj の翻訳の質が悪い、などの原因が考えられる。この調査と改善、また、対訳文を増やすことは将来の課題となる。この対訳テキストは本科研費の予算を用いて作成した。

(5) 同期 HPSG のモデル化と開発の研究成果
同期 HPSG の文法規則を開発し、(4)で開発した対訳コーパス、GIZA++による単語アライメント、英語 HPSG 構文木から、同期 HPSG 文法を獲得した。ここでは同期 HPSG 文法の句構造対応性能を評価するため、単語対に関してはテストコーパスから GIZA++によって与えられることとする。テストコーパスには(4)と同じ 1,989 文を用いた。表 5 は実験結果を示す。

表 5: 同期 HPSG 文法の性能評価

	NIST	BLEU
wsj	6.921	0.2145

実験結果をみると、得られたスコアが高いことから構造的対応がある程度とれていることがわかる。表 4 の wsj に対する Moses のスコアと比較すると、Moses よりも高いスコアを得ていることがわかるが、これはテストコーパスから単語対を得ているためと考えられる。実際の機械翻訳では、単語訳を自動的に出力する必要があるため、単語訳のモデルを必要とするが、これについては将来の課題とする。また、解析結果を調べると、GIZA++による単語アライメントが失敗している場合が多く、また、構文木の出力に失敗している場合も少なからずあった。これらの問題の解決により、スコアがさらに改善されることが期待される。同期文法の応用として同期構文解析によるパラレルコーパスの解析が期待されるが、交差などの難しい条件があるためこれについては将来の課題とする。

５．主な発表論文等

〔雑誌論文〕（計　8 件）

① Takashi Ninomiya, Takuya Matsuzaki, Nobuyuki Shimizu, Hiroshi Nakagawa: Deterministic shift-reduce parsing for unification-based grammars, Natural Language Engineering, Cambridge University Press, 査読有り, 2011 (採録決定).

② Shin Matsushima, Nobuyuki Shimizu, Kazuhiro Yoshida, Takashi Ninomiya and Hiroshi Nakagawa: Exact Passive-Aggressive Algorithm for Multiclass Classification Using Support Class, In the Proceedings of the Tenth SIAM International Conference on Data Mining (SDM10), 査読有り, 2010, pp. 303-314.

③ 松島 慎, 佐藤 一誠, 二宮 崇, 中川 裕志: PA アルゴリズムにおけるラベルなしデータの利用, 日本データベース学会論文誌, 査読有り, vol. 9, no. 1, 2010, pp. 82-87.

④ 松島 慎, 清水 伸幸, 二宮 崇, 中川 裕志: 多クラス識別問題における Passive-Aggressive アルゴリズムの効率的厳密解法, 電子情報通信学会論文誌 D, 査読有り, vol. J93-D, no. 6, 2010, pp. 724-732.

⑤ Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya and Hiroshi Nakagawa: Personalized Reading Support for Second-language Web documents by Collective Intelligence, In the Proceedings of the 2010 International Conference on Intelligent User Interfaces (IUI2010), 査読有り, 2010, pp. 51—60.

⑥ Takashi Ninomiya, Takuya Matsuzaki, Nobuyuki Shimizu and Hiroshi Nakagawa. Deterministic shift-reduce parsing for unification-based grammars by using default unification. In the Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), 2009, 査読有り, pp. 603-611.

⑦ Takashi Ninomiya, Yoshimasa Tsuruoka, Yusuke Miyao, Kenjiro Taura and Jun'ichi Tsujii. Fast and scalable HPSG parsing. Journal of Traitement Automatique des Langues (TAL), 査読有り, vol. 46, no. 2, 2007, pp. 91-114.

⑧ Takashi Ninomiya, Takuya Matsuzaki, Yusuke Miyao and Jun'ichi Tsujii. (2007). A log-linear model with an n-gram reference distribution for accurate HPSG parsing. In Proc. of IWPT 2007, 査読有り, 2007, pp. 60-68.

〔学会発表〕（計 14 件）

① 黒澤 雅人, 佐藤 一誠, 松島 慎, 二宮 崇, 中川 裕志. HMM におけるアンサンブル学習. NLP若手の会 第5回シンポジウム(奨励賞受賞), 2010

② 松島 慎, 清水 伸幸, 吉田 和弘, 二宮 崇, 中川 裕志. 多クラス識別問題におけるオンライン学習のための厳密な PA アルゴリズム.情報処理学会創立 50 周年記念（第 72 回）全国大会講演論文集, vol. 2, pp. 467-468 (学生奨励賞を受賞), 2010

③ 江原遥, 二宮崇, 清水伸幸, 中川裕志. en.newikipedia.org : 英語版 Wikipedia 中のユーザが知らない英単語を予測するユーザ参加型読解支援システム. 情報処理学会研究報告 Vol.2010-HCI-138 No.4.（2010 年度情報処理学会 HCI 研究会学生奨励賞受賞）.

④ 江原 遥, 二宮 崇, 中川 裕志. 語義注釈システムの単語クリックログからの言語能力情報の抽出. NLP若手の会 第4回シンポジウム (奨励賞受賞), 2009.

〔図書〕（計　4 件）

① Takashi Ninomiya, Takuya Matsuzaki, Yusuke Miyao, Yoshimasa Tsuruoka and Jun'ichi Tsujii. HPSG Parsing with a Supertagger. In Harry Bunt, Paola Merlo and Joakim Nivre (eds.), Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing, 2010, pp. 243--256, Springer-Verlag.

〔産業財産権〕

○出願状況（計　0 件）

○取得状況（計 0 件）

〔その他〕

二宮 崇, 宮尾祐介. 自然言語処理における文法開発の軌跡と展望. 言語処理学会第13回年次大会チュートリアル講演, 2007.

６．研究組織

(1)研究代表者

二宮 崇（NINOMIYA TAKASHI）

愛媛大学・大学院理工学研究科・准教授

研究者番号：20444094

(2)研究分担者

なし

(3)連携研究者

なし