戦略的創造研究推進事業 CREST

研究領域「高度メディア社会の生活情報技術」

研究課題「情報のモビリティを高めるための基盤技術」

# 研究終了報告書

研究期間：平成 12 年 11 月～平成 17 年 10 月

研究代表者：辻井 潤一

（東京大学大学院 情報学環 教授）

# １　研究実施の概要

## １．１　プロジェクト概要

本研究は、ネットワーク中の膨大なテキスト情報を効率的に収集し、ユーザが真に必要とする情報をわかりやすい形で提示するシステムを構築するために、言語処理と知識処理、ネットワーク・クローラーや知的エージェントの研究など、複数分野の研究を有機的に統合した基盤技術を確立することを目的とした。

５年間の研究の結果、(1) 深い意味構造を出力する文解析器、(2)テキスト構造と意味に基づく知的な索引構造、(3) 機械学習と記号処理の融合のためのアルゴリズム、(4) 意味処理・知識処理研究のリソース構築、(5) テキストの収集と処理のためのソフトウェア基盤、(6) ユーザ指向のＨＣＩ、の分野において著しい成果をあげた。

また、これらの要素的な研究成果の統合により、構造的に複雑な言語処理・知識処理を大規模なテキストベースに対して適用できる基盤技術を開発し、実際にその有効性を実験により確認した。研究の最終的な成果は、性質の異なるユーザ集団を対象とした複数の統合的サービスシステムの形態で、ユーザに公開されている。

## １．２　研究の背景と目的

大量テキストからユーザが必要とする情報を含むテキストの部分を同定する質問応答(Q／A)システムの研究は、米国を中心に活発に研究されてきた。また、質問の答えを含む複数のテキストから冗長部分を取り除き、整合性のある一つのテキストとして提示する技術も、情報融合(Information Fusion)の技術として研究されはじめている。

しかし、米国・DARPA が進めるこれらの研究は、システム構築を急ぐあまり、アドホックな技術の集積となり、見通しのよい、しかも、異なる主題分野へと移行可能な Generic な基盤技術の研究とはなっていない。たとえば、研究用プロトタイプとして作成されているシステムの大部分は、Google など既存の検索エンジンを用いてまず対象テキストを限定し、第２段階で言語処理技術を用いた処理を行う方式をとる。このため、対象テキストを限定する第一段階には、言語処理、知識処理の成果は全く使えない。また、第２段階の言語処理も、実時間処理の時間的な制約から本格的な技術は使用できず、パターン照合など非常に単純な処理が使われている。知識に基づく推論処理、深い意味に基づく言語処理の重要性は強調されているが、実態は、初歩的で単純な技術を大量テキストに適用することにとどまっている。

これらの欠陥は、

1. 短期的な成果を求めて既存の検索エンジンを使用するために、テキストの収集と索引構造がBlack Boxとなり、知的処理を収集や索引構造に反映できない
2. 一般分野のQ/Aという、実ユーザの要求が不明確なタスク設定のために、分野固有の詳細な知識を活用できず、架空性が強い研究になっている

という、研究枠組み自体の不備と米国の短期的なファンデングの形態によるものである。

これに対して、本プロジェクトでは、５年間という長期的なプロジェクト設計のもとに、言語処理、知識処理、ソフトウェア技術、エージェント技術という、異なった分野

のきっちりとした研究成果を積み重ねることで、最先端の関連技術を有機的に統合した基盤技術を開発することを目標とした。

このような複数分野の最新成果をテキスト情報の収集・処理・提供のために統合して、系統的な基盤技術を開発する研究は、CREST 研究発足時の５年前には、世界でも全く行われていないものであった。

## １．３　研究項目と成果

プロジェクトは、理論と要素技術の研究に重点をおく前半３年間と、それらを統合することで技術の有効性を実証する後半２年間に大別し、メリハリのある遂行を目指した。前期の要素技術の研究は、(i) 文解析と情報抽出を中心とするテキスト処理技術、(ii) 特定分野のオントロジー構築と知識処理、(iii) 膨大なテキストの収集と処理を支える基盤ソフトウェア技術、(iv)ユーザ指向の情報交換を支えるエージェント技術、の４つの領域で研究・開発を行なった。

以下に、要素技術の主な研究成果を示す。

**(1) HPSG による英文解析システム（Enju）［３－１節］**：言語学的に妥当な文法に対する確率モデルの定義に成功し、制約を満す結果をすべて出力する能力を持つだけでなく、コーパスから獲得された確率モデルに従って、おのおの結果の確率値も同時に付与することができる。また、コーパス指向の文法開発を提唱し、Generic な言語モデルを現実のコーパスに合わせて精緻化する技術、分野固有の文法制約を学習する手法を確立した。これらの理論を実装した文解析システム Enju は、深層の意味構造を計算する Deep Parser で、かつ、大規模テキストを現実的な時間で解析できる世界最初のものである（３－７節、統合実験を参照)。文法の精緻化により、精度の点でも、現在、世界有数のものとなっている（表 1.3.1)。

［表 1.3.1］　英文解析システム Enju の解析精度

| | 適合率 | 再現率 |
|---|---|---|
| ベースライン | 76.9% | 76.8% |
| 既存のモデル | 80.6% | 81.0% |
| 提案モデル(Feature forest) | 88.0% | 87.2% |

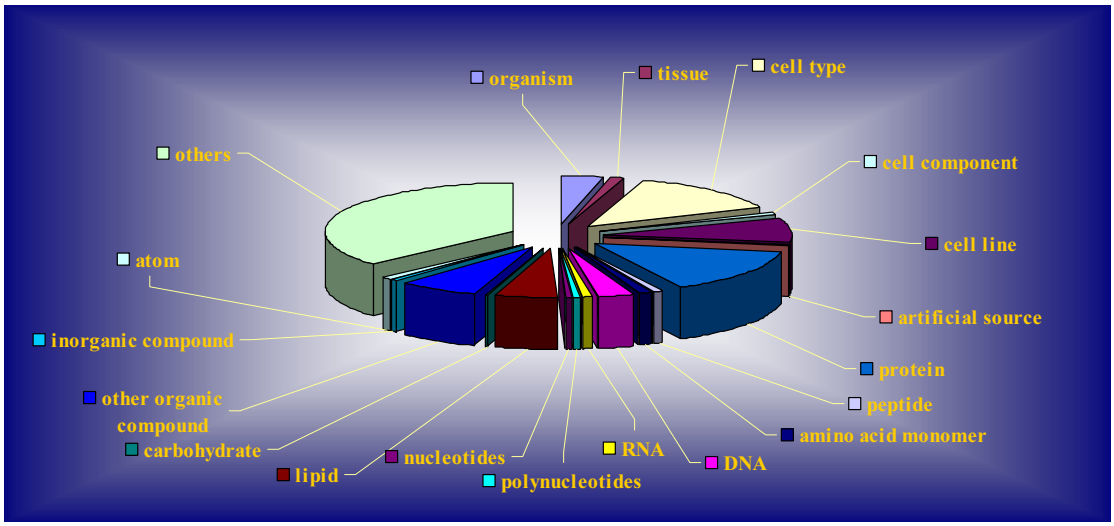
［実験環境］
Penn Treebank（Wall Street Journal）Section 02-21 で学習したモデルを用いて、Section 23 に対する精度（述語項関係の適合率・再現率）を測定。

**(2)高効率な領域代数処理システム、素性構造データベースの開発［３－２節］**：埋め込み型のタグをもつ XML データに対しても高効率で検索を行う領域代数システムを開発し、（１）の処理結果を索引構造に反映させることで、指定の意味構造検索を高速に検索するシステムを開発した。また、これを使った統合サービスシステムとして、１４億語の解析ずみ Medline での検索システムを構築し、現在、EBI(European Bioinformtics Institute)で運用されている同種のシステムより優れた検索能力を持つこ

とを確認した。

**(3) 分野に依存した POS タガー、NER の開発［３－３節］**：MEMM の新しい最適化手法として、双方向・最尤探索の手法を提案し、それが CRF(Conditional Random Field)と同様に大域的な最適化効果をもち、かつ、はるかに高速な実装が可能なこと、また、英語品詞付与や NER に有効であることを実証した。さらに、この同じ手法を英語の POS タガーや生命科学の NER に適用して、従来手法よりも優れた精度がでることを確認した。これらの成果は、個別のタスクごとに工夫されてきた機械学習と処理アルゴリズムの関係をタスクに独立な一般的な枠組みで統合するものであり、国際的にも高い評価を得ている。

**(4)生命科学の意味つきコーパスとオントロジー（GENIA コーパス）［３－４節］**：２０００抄録（２万文、５０万語）の AnnotatedText（GENIA コーパス）を作成し、世界に公開（図 1.3.1）した。現在、２４０を越える研究チームがこのデータを使って研究している。また、このデータは、いくつかの国際ワークショップでのゴールド・スタンダードとしても活用されている。GENIA コーパスは、意味タグのついた世界最大規模のコーパスであるだけではなく、文の構文構造・名詞句間の参照関係・生命現象にかかわる事象など、非常に豊かな情報が付与されていることから、急速に展開しつつある言語的意味と知識の研究においても、重要な基礎資料となる。




［図 1.3.1］ GENIA コーパス中での意味クラスの分布

**(5)テキスト収集の高速クローラーと分散処理のための基盤ソフトウェア GPX［３－５節］**: 高速クローラーの研究では、ひとつのウェブサーバへのアクセスを集中的に行うことなく、1 台の計算機で 600 ページ/秒のダウンロード速度を数時間維持することに成功した。また、その crawler を 20 台程度で並列に実行し、この台数まで性能がスケーラブルに向上する(約 10,000 ページ/秒)ことを確認した。また、日常的に利用できるＰＣクラスタを簡便に使用するためのツール GPX を開発し、実際のテキスト処理に適用することでその有効性を確認した（３－７節、統合実験を参照)。

**(6) ユーザ指向のＨＣＩ［３－６節］**：言語的・テキスト的な媒体による意思の伝達の対極として、身体性を考慮した情報伝達を、人間とロボットとのインタラクションを例に研究した。とくに、引き込み原理と確率推論を使った人間・人工物インタラクション方式を開発し、それを実際のロボットに実装することで、その有効性を確認した。引き込み原理を使ったインタラクション方式は、個々のジェスチャの意味づけを予め行わない非記号的なジェスチャによるインタラクションを、世界に先駆けて実現したものである。

プロジェクト後半の２年間は、これらの要素技術だけでなく、それらを統合する研究も積極的に推進し、統合的な実験や統合サービスシステムの構築を行った。以下の統合実験は、我々の技術が国際的な研究水準をはるかに凌駕するものであることを実証した。また、多様な実ユーザを対象にしたサービスシステムを構築することで、研究目的の項（１－２節）で述べた「Generic な言語処理に基づくアドッホクでない基盤技術で、かつ、実ユーザの情報要求にこたえるための基盤技術」が可能なことを実証した。

**(7)統合実験［３－７節］**：分散計算環境のツール GPX、HPSG による文解析器(Enju)、機械学習による POS/NER、ＧＥＮＩＡコーパスという、４つの研究成果を統合することにより、生命科学分野のテキストベース Medline 抄録全体（１４億語）を、８日間で処理することに成功した（表 1.3.2)。これは、深い文解析を使った実験としては、従来の研究を質・量ともに大きく凌駕するもので、この種のシステムでは最速の Enju を使っても、分散処理を行わない場合には、２年以上の時間を要する。

［表 1. 3. 2］ 並列分散処理による文解析実験の結果

| |
|---|
| [実行環境]<br>PC クラスタ （合計 350 プロセッサ）<br>IBM BladeCenter Cluster Dual Xeon 2.4GHz × 69 ＋ Dual Xeon 2.8GHz × 42<br>Appro Blade Server Dual Xeon 2.4GHz × 64<br>並列実行プラットフォーム： GXP (Grid Explorer) |
| [実験データ]<br>MEDLINE 1500 万論文<br>規模： 7000 万文（14 億語）<br>処理時間：８日間 |

**(8)統合的なサービスシステム［３－８節］**：明確な情報要求を持ったユーザ（生命科学分野の研究者）と緊密な共同研究を行ない、本プロジェクトの成果をかれらのためのサービスシステムとして実現することで、研究の有効性を示した。具体的には、病疾患・遺伝子の関係発見を援助するシステム、蛋白質相互作用の抽出システム、Medline の知的検索システムを作成した。また、ユーザを特定しない専門用語認識システム(言選)や多言語情報検索システム(KIWI)を開発、公開した。

## ２　研究構想及び実施体制

### ２．１ 研究構想

#### (1) 研究の方向、考え方

インターネット・Ｗｅｂに代表される計算機ネットワークの中に蓄積・流通するテキスト情報を有効に活用するためには、言語処理、オントロジーに基づく知識処理、ソフトウェア技術という、かなり性質の異なる研究を有機的に組織化し、系統的な研究・開発を行う必要がある。一見、面白く見える研究も、それをスケール・アップする堅固な技術基盤がないと、Research Prototype という、絵に描いたもちに終わる。

このことから、５年間という比較的長い研究期間を考え、スケール・アップに耐える要素技術やリソースの構築に焦点を当てる前期３年間と、より統合的でユーザ指向的なシステム開発に重点をおく後期２年間に分け、めりはりのある研究遂行を目指した。

とくに、前期は、研究を特定応用のためのアドホックな技術開発にしないために、（１）理論的に健全な枠組みに基づく汎用の言語処理技術、（２）本格的なオントロジー処理のための大規模リソースの構築、（３）大量、かつ、複雑な処理を行うための計算機インフラ技術、（４）ユーザ指向の情報交換を目指すエージェント技術、を目指し、４つのグループ（言語処理Ｇ、オントロジーＧ、ソフトウェアＧ、エージェントＧ）がそれぞれを担当した。また、明確な情報要求をもつユーザ集団（生命科学の研究者）と緊密な研究協力を行うことで、後期には、現実的な場面で実際に有効に使用できる統合システムを開発することを心がけた。

#### (2) 推進体制

**[研究グループ内での協力]**

実際の研究推進では、個々の技術分野での研究開発を強い連携のもとに行うために、前期には平均月一回の研究交流と打ち合わせの会を持ち、後期には、グループ間の研究成果を実際に統合した研究を行った。その結果、ソフトウェアＧのクローラーが収集したテキスト集合を使うことで、言語処理 G やオントロジーG が言語学習の実験、索引構造の構築を行った。さらに、統合実験において、生命科学のテキストベース Medline（１５００万抄録、１４億語）のすべてに対して、意味構造を付与する解析実験を行い、その結果を世界的に公開できたのは、言語Ｇ、ソフトウェアＧ、オントロジーＧの緊密な協力の結果であった。

**[外部グループとの協調]**

実ユーザの要求に応えるための基盤技術構築を目指す本プロジェクトでは、実ユーザとの共同研究が不可欠である。研究開始の当初から、生命科学への応用をひとつの柱としていたことから、産総研・JBIRC、国立情報学研究所、特定研究「ゲノムサイエンス」の研究者と緊密な連携のもとに研究を進めた。特に、JBIRC が推進する H-INV のプロジェクトに Non-Funded Partner として参加した。統合サービスシステムの病疾患・遺伝子の関係発見システムは、かれらとの共同研究の成果である。

**［国際協力］**

本プロジェクトでは、国際的な研究をリードすることに留意し、国際ワークショップを開催するなど、われわれの研究成果を積極的に宣伝した。また、有力な海外グループ(ペンシルベニア大、スタンフォード大、米国・Parc、マンチェスター大、EBI、ＣＮＲＳなど)と積極的な研究協力を推進し、研究の実質面での協力関係を確立した。たとえば、Coling(２００４)での GENIA を中心とするワークショップは、財政的には本プロジェクトとは独立であるが、本プロジェクトの JST 雇用研究員が大きな役割を果たしており、GENIA コーパスが実質的な世界標準となる契機を作った。

また、中国・海南島で開催した「深い言語解析」についての WS には、ペンシルベニア大学・ダブリン大学・ケンブリッジ大学など、この分野の有力研究グループの参加を得ることで、この方向での研究が国際的に活発化するもととなった。その結果、我々の基本技術（素性構造文法の確率モデルと素性森、GENIA タガー、英文解析器）が多くの海外グループの研究で引用され、実際に使われている。

**(3) 新たな展開**

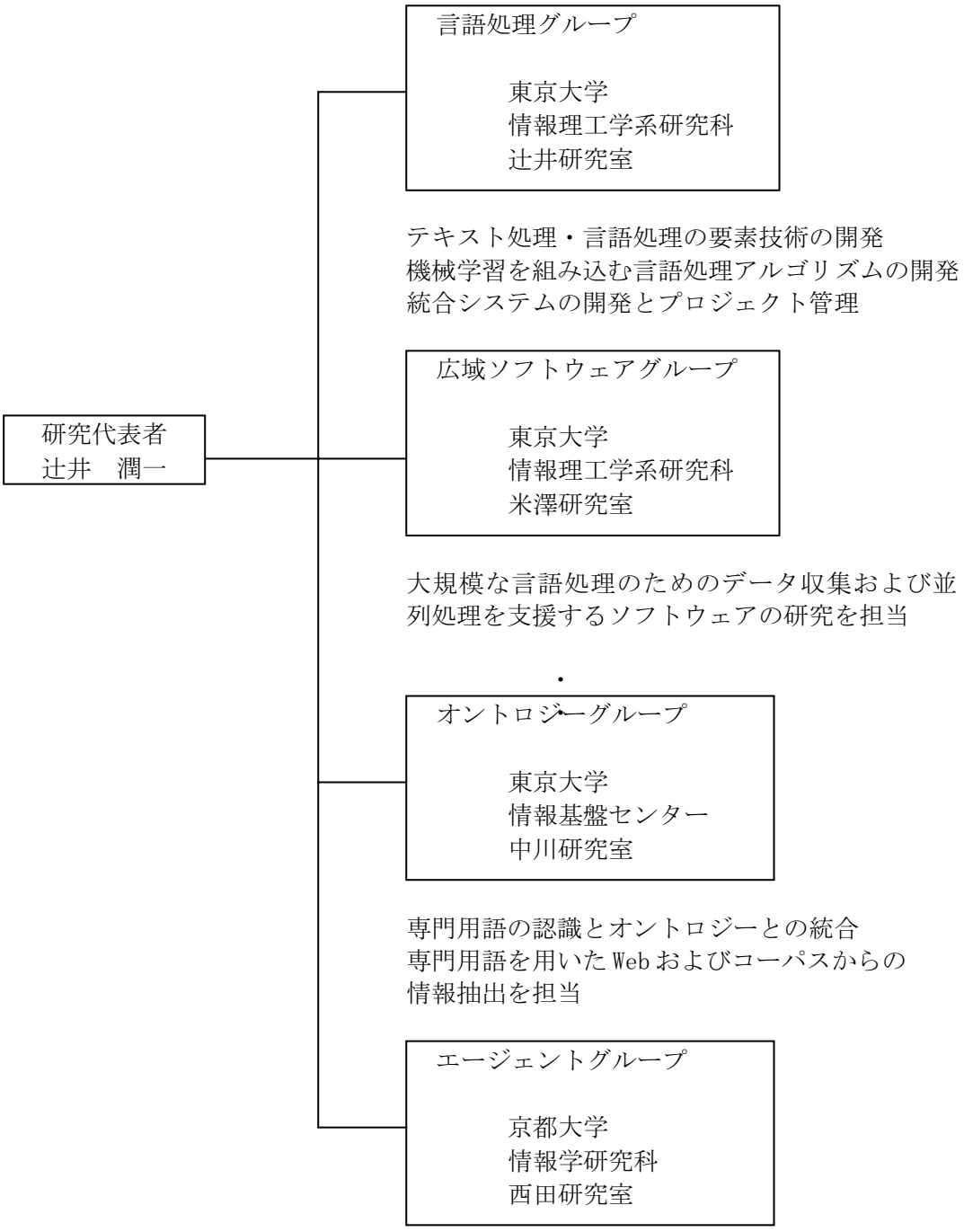
我々が目指した方向（スケーラブルな言語処理、情報抽出における深い意味処理の活用、生命科学でのテキストマイニング、など）は、５年前には野心的すぎる、需要が顕在化していないなど、批判と危惧があったが、この５年間に、そのいずれもが大きな潮流となり、それぞれをテーマにした国際会議、ジャーナル特集号が企画されるようになっている。LREC、Coling、ACL、HLT、SMBM など、分野の有力な国際会議での基調講演、チュートリアル講演にグループの研究者が招待されるなど、我々の研究グループは、これらの新しい潮流を主導する立場になっている。

たとえば、スケーラブルな言語処理は、膨大なテキストの集積と GRID 技術により、過去５年間に急速にその可能性と有効性が議論されるようになり、そのための研究センター（英国：マンチェスター大学、ドイツ：イエナ大学、米国：コロラド大）が設立されている、あるいは、されようとしている。

また、生命科学のテキストマイニングでは、言語処理の国際会議(Coling、ACL など)や生命科学の国際会議(ISMB、PSB、ICSB など)に特別セッションやワークショップが企画された。

情報抽出・情報検索への深い言語処理の適用は、必要な言語処理技術をもつグループが少ないため、我々以外では、ケンブリッジ大・ザール大学・エディンバラ大・Parc・ペンシルベニア大・スタンフォード大などに限られている。ただ、これらは、計算言語学を世界的にリードしてきたグループであり、今後、この方向への研究が加速されると予想される。浅い処理という、アドホックな技術の時代が終焉し、きっちりした基盤をもつ技術の開発期に入った。この潮流の変化をもたらした一つが、われわれが組織した、前述のワークショップ(2004 年、中国・海南島)であった。

## ２．２ 実施体制


研究代表者
辻井　潤一
言語処理グループ
東京大学
情報理工学系研究科
辻井研究室
テキスト処理・言語処理の要素技術の開発
機械学習を組み込む言語処理アルゴリズムの開発
統合システムの開発とプロジェクト管理
広域ソフトウェアグループ
東京大学
情報理工学系研究科
米澤研究室
大規模な言語処理のためのデータ収集および並
列処理を支援するソフトウェアの研究を担当
オントロジーグループ
東京大学
情報基盤センター
中川研究室
専門用語の認識とオントロジーとの統合
専門用語を用いたWebおよびコーパスからの
情報抽出を担当
エージェントグループ
京都大学
情報学研究科
西田研究室
ユーザ指向の情報と提示と情報交換の研究

## ３　研究実施内容及び成果

### ３．１　高効率・高耐性・高精度な文解析システム(Enju)の開発
### ［東京大学　言語処理グループ］

#### (1) 研究実施内容及び成果

高度な情報抽出・情報検索を実現するための基盤技術の一つに、単語列から意味構造を計算する文解析の技術がある。とくに単なる構文構造だけでなく、より深い意味構造までの計算を系統的に行うには、言語学的な理論に整合的で、かつ、現実のテキストに現れる幅の広い言語現象に対応できる高耐性な文法モデルの開発が不可欠となる。このような理想的な文法モデルが開発できなかったために、工学的な応用を重視する研究では、特定の分野と応用に特化した、アドホックな技術が使われてきたといえる。

本プロジェクトでは、（１）現実のコーパスをもとに文法を作成すること、ただし、作成される文法は、言語理論に整合的なものであり、かつ、その文法は、特定の分野用に Tuning できる Generic なものであること、また、（２）その文法による文解析が、現実の大量テキストの処理に使える高い処理効率をもつこと、を目指して研究を行ってきた。以下では、文法モデルの構築とその高速処理の２項目に分けて、報告する。

#### ① Enju の文法モデルと文法開発、性能の評価

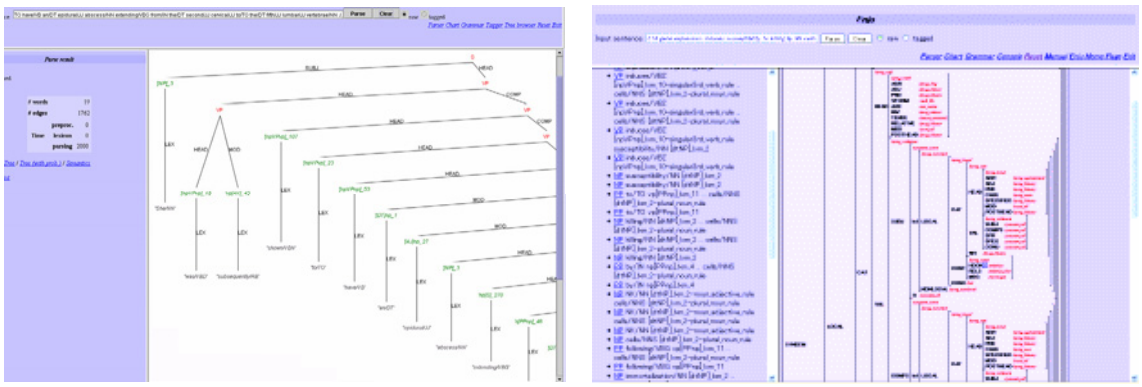
言語理論に基づく詳細な構文構造の解析は、自然言語の意味を計算するために必要不可欠な技術として、長期間にわたり研究されてきた。言語理論では、人間が意味を文章として表出する際、様々な構文的変形（受身、倒置、関係詞節など）を受けて実際の文が出現すると考える。すなわち、構文構造が分かれば、どのような構文的変形が起きたかが分かるため、文の表現する意味を計算することが可能となる。

理論的研究としては、主辞駆動句構造文法(Head-Driven Phrase Structure Grammar, HPSG)や語彙機能文法（Lexical Functional Grammar, LFG）など、語彙化文法理論とよばれる言語理論が広く研究され、一定の成功を収めている。それに伴い、これらの文法理論を実際にコンピュータプログラムとして実装し、実世界のテキストの構文・意味構造を自動的に計算する試みも盛んに行われている。しかし、現在のところ、新聞や論文などの自然言語テキストを満足に解析できる構文解析システムはほとんど存在しない。実世界のテキストを解析するために利用されている既存の解析器は、言語学的な理論づけがなく、表層的な木構造や単純な係り受け構造を出力するにとどまっている。

本研究では、詳細な構文・意味構造を出力しかつ実用的な構文解析器を開発するため、２つの問題点に着目し、それらを解決する手法を提案した。一つは、大規模な文法を効率的に開発するための方法論、もう一つは、複雑なデータ構造で表現される構文構造に対して確率モデルを推定するためのアルゴリズムである。

これらの手法により、大規模な英文解析システム Enju を開発することに成功した（図 3.1.1）。Enju は HPSG に基づく構文構造、さらには深層の意味構造（述語項構造）を自動的に計算する構文解析器である。新聞のテキスト(Wall Street Journal)を用いた評価実験の結果、99.5%の文に対して構文・意味構造を出力することができ、また、88%

程度の精度で単語間の意味的関係（述語項関係）を出力することを実証した。このように詳細な構文・意味構造を高い精度で計算できる構文解析器は、世界有数のものとなっている。



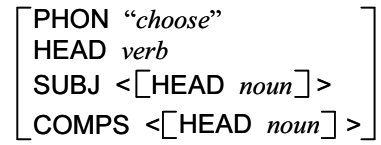
［図 3.1.1］英文解析システム Enju

さらに、新聞のテキストでトレーニングした構文解析器を生命科学論文のテキストに適応させる研究を行った。Enju は汎用の英文解析器であるが、曖昧性解消のための確率モデルを GENIA Treebank を学習データとして再トレーニングさせた結果、少ない学習データを用いた場合でも高精度を達成することを実証した。この構文解析器を用いて、Medline 全体（１５００万抄録、１４億語）の英文テキストの構文解析を行い、その結果が検索システム等で利用されている。詳細な構文・意味構造を出力し、かつこのような大規模データを処理できる構文解析器は、世界で最初のものである。

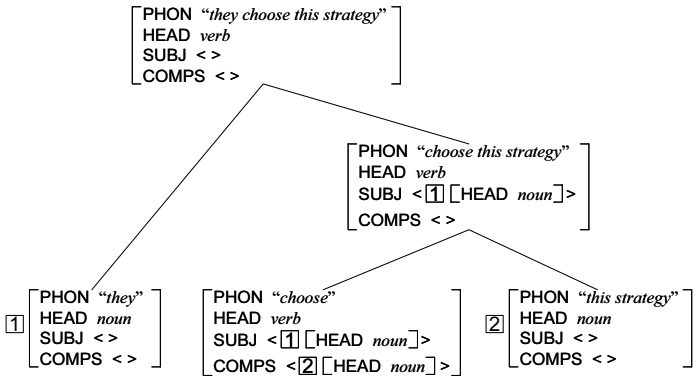
以下では、文法開発と確率モデル及び評価実験について、詳細を報告する。

**コーパス指向の文法開発**

HPSG は、言語の一般的規則性を規定する*プリンシプル*と、単語固有の性質（品詞や下位範疇化フレームなど）を表現する*語彙項目*との組み合わせで、自然言語の構文・意味構造を説明する文法理論である。例えば、図 3.1.2 に他動詞“*choose*”の語彙項目を示す。“HEAD”は単語の品詞を表し、この場合は動詞なので *verb* と記述されている。“SUBJ”、“COMPS”はこの単語が主語、目的語としてとりうる単語が持たなければならない構造を表す。“*choose*”は他動詞なので、それぞれ一つずつ名詞（つまり“HEAD”が noun である）をとらなければならないと記述されている。実際の語彙項目には、格、性数、時制や意味構造等に関する詳細な制約が書き込まれ、一つの語彙項目に書き込まれる制約は 100 以上に及ぶこともある。HPSG に基づく構文解析は、文中の単語の語彙項目が持つ制約を満たすように構文木を組み立てることで行われる。例えば、図 3.1.3 に示すように、“*they choose this strategy.*”という文を解析すると、“*choose*”は主語として“*they*”、目的語として“*this strategy*”をとることができるため、文全体の構文構造が完成する。図 3.1.3 において、[1]や[2]は変数を表し、同じ数字の構造は同じ制約を持たなければならないことを表している。例えば、[1]は“*they*”の語彙項目に記述されている制約と“*choose*”の主語の制約が一致しなければならないことを表している。




［図 3.1.2］“*choose*”の語彙項目




［図 3.1.3］HPSG に基づく構文解析

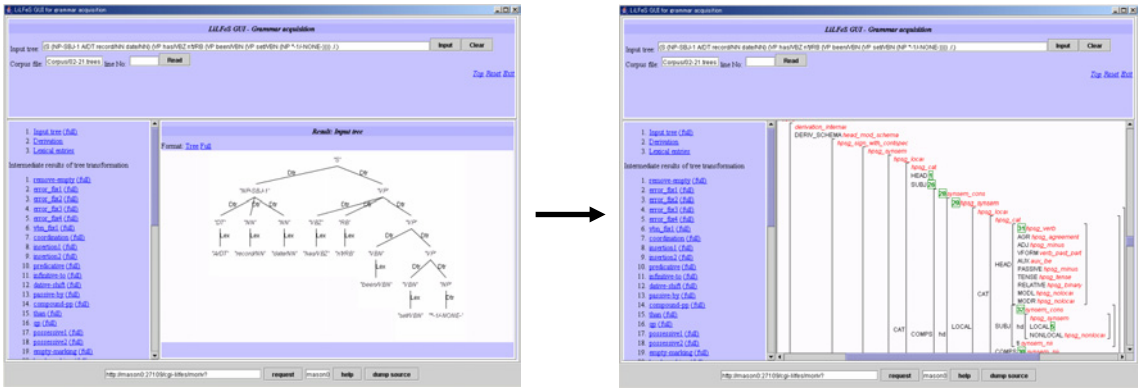
HPSG に基づく構文構造が計算できれば、文の表現する意味構造が得られるため、コンピュータに実装するために様々な研究が行われてきた。しかしながら、実世界のテキストを解析できるまでの大規模な文法を開発するのはほぼ不可能であった。その主な原因は、文法のスケーラビリティの問題であったと考えられる。HPSG に基づく文法で実世界のテキストを解析するためには、大量の単語の様々で詳細な構文的振る舞いを語彙項目に記述する必要がある。すると、新たな単語や構文構造を説明するために文法を改変すると、その改変が他の語彙項目やプリンシプルとの関係に影響を与え、文法の一貫性が保てなくなる。具体的には、ある単語・構文を新たに扱うために文法を改変すると、その改変により、それまで解析可能であった文が解析できなくなるといったことが頻繁に発生する。語彙項目やプリンシプルはほぼ無限の組み合わせ方が可能性であり、そのために様々な文を生成することができるわけであるが、その全ての組み合わせの可能性を考慮しつつ文法を改良していくことは通常の人間には不可能である。

本研究では、文法のスケーラビリティの問題を解決するために、文法開発の新しい方法論として**コーパス指向の文法開発**(Miyao et al., 2003a, 2005; Nakanishi et al., 2004a, 2004b)を提案した。この手法では、語彙項目を開発する代わりに*ツリーバンク*を作成する。ツリーバンクとは、実世界の文に対して、その文の構文解析木を人手で与えたデータである。すなわち、構文解析の正解例のデータであり、構文構造の実例のデータである。十分な大きさのツリーバンクを開発することができれば、図 3.1.3 に示すように解析木の終端ノードが語彙項目であるため、ツリーバンク中の解析木の終端ノードを集めることで語彙項目が自動的に得られる。

本プロジェクトでは、構文解析の研究で広く利用されている Penn Treebank を、HPSG ツリーバンクに変換することにより、大規模な英語 HPSG 文法を開発した（図 3.1.4）。Penn Treebank は新聞（Wall Street Journal）のテキストに対して人手で表層的な木構造の情報を与えたコーパスである。これを変換することにより、HPSG 理論に基づき詳細な構文構造が書き込まれた HPSG ツリーバンクを得ることに成功した。

まず、文法開発者は HPSG の理論に則ったプリンシプルを定義する。次に、既存の言語リソース（Penn Treebank）を、HPSG 理論に則ったツリーバンク（HPSG ツリーバンク）に変換する。文法開発者の主な仕事は、ツリーバンクの変換過程をコントロールすることとなる。その際、HPSG ツリーバンク中のエラーや非一貫性は「プリンシプルの

違反」という形で自動的に検出することができる。つまり、文法開発者はエラーや非一貫性の原因を簡単に同定することができ、それを修正していくことでより高品質なツリーバンクを作り上げていくことができる。すなわち、本方法論における文法開発とは、文法理論に則ったプリンシプルを満たすようにツリーバンクを作成していくことを意味する。本手法により、文法の一貫性の制御が容易になり、また既存の言語リソースが再利用できるため、文法開発のコストが大幅に抑えられる。



［図 3.1.4］Penn Treebank から HPSG ツリーバンクへの変換

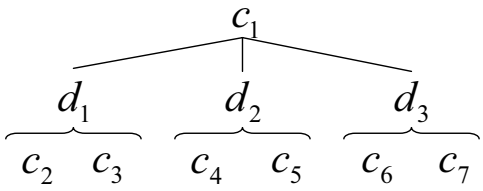
**確率モデルによる曖昧性解消**

実用的な構文解析には曖昧性解消が必須である。HPSG 文法は文法的に正しい構文構造を導出するように設計されるが、一般的に自然言語の文は文法的に正しい構造を複数持ちうる。アプリケーションは通常、曖昧性が解消された解析結果を要求する。したがって、構文解析器は全ての解候補の中から最適な解析結果を出力することが要求される。

自然言語処理では、確率モデルが曖昧性解消のために広く用いられ、成功を収めている。従って、我々の問題に対しても確率モデルが適用可能と考えられる。しかしながら、HPSG 文法は図 3.1.2、図 3.1.3 に示したような複雑なデータ構造（型付き素性構造）で記述されているため、既存の確率モデルは適用できない。既存の自然言語処理技術は、文全体の構造を部分構造に分解し、その部分構造間の統計的独立性を仮定し、部分構造の確率モデルを推定する、というアプローチをとる。例えば、文の品詞列は単語ごとの品詞、文全体の構文木は一段ごとの分岐に分解され、分解された部分構造の確率を推定し、全体構造の確率は部分構造の確率の積として定義する。これらの手法は、対象問題の構造が木構造や格子構造であることを前提としている。HPSG などで用いられる型付き素性構造は任意のグラフ構造であるため、この手法は適用できない。一般に、型付き素性構造の確率モデルを従来手法で推定すると、最適でない確率モデルが得られることが知られている。

この問題に対する解決策として、本研究では **feature forest モデル**(Miyao and Tsujii,2002)を提案した。Feature forest とは、指数関数的数の木構造をパックした構造で表現する一般的データ構造である。例えば、図 3.1.5 において、ノード $d_i$ は曖昧性を表し、その子ノードうちどれか一つが選択されることを表す。従って、図 3.1.5 の feature

forest は２×２×２＝８個の木構造を表現している。Feature forest モデルは、feature forest の上に定義された最大エントロピーモデルである。本研究では、feature forest を展開することなく最大エントロピーモデルのパラメータを推定する動的計画アルゴリズムを提案した。これにより、feature forest のノードがどのようなデータ（型付き素性構造など）であっても、統計的独立性の仮定なしに最尤な確率モデルを推定することができる。したがって、複雑なデータ構造で表現された事象の確率モデルを構築することが可能となった。




［図 3.1.5］feature forest

さらに、本研究では、HPSG の構文解析木や述語項構造が feature forest で表現できることを示した(Miyao and Tsujii, 2003, 2005; Miyao et al., 2003b)。これらにより、HPSG に基づく構文解析のための確率モデルを構築する方法を確立した。

また、確率モデルを一から学習するのではなく、既存の確率モデルを特定分野のテキストに適応させる手法を提案した(Hara et al., 2005)。一般に、汎用の構文解析器を特定分野のテキスト（生命科学論文など）にそのまま適用すると、解析精度が低下することが知られている。しかし、各分野ごとに大規模な学習データを作成するのは現実的でない。そこで、オリジナルの確率モデルの情報を残したまま、特定分野に特化させるアルゴリズムを提案した。本手法は、分野とタスクに依存した構文解析システムを構築するための基盤技術となる。

**HPSG に基づく英文解析器の実装と評価実験**

本研究で提案した手法を実装し、HPSG に基づく英語構文解析器 Enju の開発を行った。英語新聞記事のツリーバンクである Penn Treebank をテストデータとして、構文解析の実験を行い、Enju の実用性および提案手法の有効性を示した。表 3.1.1 に Enju 文法の規模と構文解析実験の結果を示す。

| | |
|---|---|
| 単語数 | 34,754 |
| 語彙項目数 | 1,942 |
| 文被覆率 | 99.5% |
| 述語項関係の解析精度<br>（適合率／再現率） | 88.0%/87.2% |
| １文平均解析時間 | 0.56 秒 |

［表 3.1.1］Enju 文法の規模と評価実験の結果

まず、Penn Treebank 約４万文を HPSG ツリーバンクに変換するプログラムを実装し、そこから約 35,000 単語に対して 2,000 近くの語彙項目を獲得した。新聞のテキストに対する文法の被覆率（解析できる文の割合）を測定した結果、99.5%の文に対して何らか

の解析結果を出力することができ、またそのうち 84.1%の文に対しては正解の構文木を含む構文解析候補を出力することを確認した。この被覆率は現在までの構文解析器では達成し得ない高い性能を示している。

次に、feature forest モデルを適用し、HPSG に基づく構文解析の曖昧性解消モデルを開発した。曖昧性解消の結果、単語間の意味的関係（述語項関係）の精度を測定したところ、88.0%/87.2%（適合率/再現率）の精度を達成した。詳細な意味構造をこのような高い精度で出力できる構文解析器は世界有数のものである。また、構文解析時間についても、様々な高速化技術を開発することにより、1 文平均 0.5 秒という解析速度を達成している（詳細は次節を参照）。

分野適応手法の有効性を検証するため、GENIA Treebank 約 4,500 文を用いて再学習した Enju の解析精度を表 3.1.2 に示す。オリジナルの Enju に比べて 1 ポイント以上の精度向上が見られ、表 1 に示した新聞のテキストに対する精度に近い精度を達成している。オリジナルの Enju の学習データと比べて 1/10 程度のデータしか用いていないが、オリジナルの確率モデルと特定分野の学習データを適切に融合することにより、少ない学習データでも高い精度を達成することを実証した。

| | |
|---|---|
| オリジナルの Enju | 85.40%/83.75% |
| GENIA Treebank のみで学習した Enju | 86.38%/79.96% |
| 再学習させた Enju | 86.82%/85.25% |

［表 3.1.2］生命科学論文に対する解析精度

本研究では、実世界のテキストを HPSG に基づき解析することに初めて成功した。実用面では、実世界のテキストに対して詳細な構文・意味構造が自動的に計算できるようになったことで、様々な自然言語処理への応用が期待される。実際、本プロジェクトでは、生命科学論文に適応した Enju を用いて Medline 抄録を解析する大規模な実験を行っている（詳細は 3.2 節を参照）。また基礎研究としては、本研究により、言語学に基づく詳細な構文解析を実用レベルのシステムとするための基本的な方法論を確立し、その有効性を実証した。

## ［関連発表文献］

(Miyao et al., 2003a) Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. (2003). **Lexicalized grammar acquisition.** In Proc. 10th EACL. pp. 127-130.

(Miyao et al., 2005) Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. (2005). **Corpus-oriented grammar development for acquiring a Head-Driven Phrase Structure Grammar from the Penn Treebank.** In Natural Language Processing – IJCNLP 2004. LNAI 3248. pp. 684-693.

(Nakanishi et al., 2004a) Hiroko Nakanishi, Yusuke Miyao and Jun'ichi Tsujii. **Using Inverse Lexical Rules to Acquire a Wide-coverage Lexicalized Grammar.** In the IJCNLP 2004 Workshop on Beyond Shallow Analyses. 2004.

(Nakanishi et al., 2004b) Hiroko Nakanishi, Yusuke Miyao and Jun'ichi Tsujii. **An Empirical Investigation of the Effect of Lexical Rules on Parsing with a Treebank Grammar.** In the Proceedings of the third TLT2004. pp. 103--114. 2004.

(Miyao and Tsujii, 2002) Yusuke Miyao and Jun’ichi Tsujii. (2002). **Maximum entropy estimation for feature forests.** In Proc. HLT 2002. pp. 292-297.

(Miyao and Tsujii, 2003) Yusuke Miyao and Jun’ichi Tsujii. (2003). **A model of syntactic disambiguation based on lexicalized grammars.** In Proc. 7th CoNLL. pp. 1-8.

(Miyao et al., 2003b) Yusuke Miyao, Takashi Ninomiya, and Jun’ichi Tsujii. (2003). **Probabilistic modeling of argument structures including non-local dependencies.** In Proc. RANLP 2003. pp. 285-291.

(Miyao and Tsujii, 2005) Yusuke Miyao and Jun’ichi Tsujii. (2005). **Probabilistic disambiguation models for wide-coverage HPSG parsing.** In Proc. ACL 2005. pp. 83-90.

(Hara et al., 2005) Tadayoshi Hara, Yusuke Miyao, and Jun'ichi Tsujii. **Adapting a probabilistic disambiguation model of an HPSG parser to a new domain.** In Robert Dale, Kam-Fai Wong, Jian Su and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2005. LNCS3651. Springer-Verlag. 2005.

### ② Enju の高速な文解析アルゴリズム

我々は①で説明した主辞駆動句構造文法 (HPSG) の構文解析器 (HPSG パーザー) Enju の高速化のために、ビームサーチによる高速化手法と、深い構文解析のための様々な高速化手法の適用を行った (Ninomiya et al. 2005)。まず CFG のためのビームサーチアルゴリズムとビームサーチの高速化と再現率向上のためのイテレィティブ構文解析 (Tsuruoka and Tsujii 2005a) を HPSG 構文解析に適用し、続いて、クイックチェック、 大句構造の抑制、および CFG チャンクパーザー (Tsuruoka and Tsujii 2005b) とのハイブリッド構文解析による高速化を実現した。構文解析の性能評価によく用いられる Penn Treebank (Wall Street Journal)に対し実験を行い、それぞれの手法を精度と解析速度の点から評価した。

単一化文法は言語学的形式化と計算効率の観点から広く研究をなされてきた。しかし、それらは非常に精緻な言語学的構造を文に対して与えるものの、その精緻な記述のために計算効率は非常に悪いとみなされてきた。高速化は実用アプリケーションのための主たる目的の一つであるため単一化文法の構文解析効率のための研究が非常に多くなされてきた。それらの研究により解析効率が大幅に改善されたものの、構文解析速度はまだ実用的なアプリケーションのためには十分ではなかった。

近年、広範な解析能力を持つ単一化文法のための確率モデルが提案され、確率値を参照することにより、より効率的な構文解析の探索を行う手法の可能性がでてきた。つまり、一般に最も確率値の高い構文木だけが解析結果として要求されるので、最終的な解析結果に貢献する可能性の高い部分解析だけ計算すればよい。この方法はビタビ構文解析やビームサーチ構文解析に代表される確率付き CFG 構文解析で広く研究されてきた。

多くの単一化文法のための確率的構文解析が開発されてきたが、それらの多くは、まず全ての構文解析結果を網羅的に求めておき、その後、もっとも高い確率の構文木を一つ選択する、という方法をとっていた。しかし、これらの手法の効率は必然的に最初の網羅的な構文解析の効率に依存してしまうという短所がある。CFG バックボーンに対して確率を付与し、確率的 CFG 構文解析により最も確率の高い構文木を選択する手法も提案されているが、これらは確率的 CFG モデルに過ぎず、確率的単一化文法モデル

ではない。

今まで、大規模コーパスに対し、単一化文法の確率モデルを用いてビームサーチ構文解析を行った例はなく、我々は大規模コーパスに対する確率的 HPSG 構文解析にビームサーチを適用する手法の考案、およびその性能について報告をする。また、深い構文解析のための種々の高速化手法の性能についても実験で確認する。

**深い構文解析のための高速化手法の適用**

**［クイックチェック］**単一化文法である HPSG は全ての文法規則の適用に単一化が行われるため、単一化の高速化は構文解析全体の高速化に大きく貢献する。クイックチェックは、単一化に失敗する素性構造の組み合わせを除去するフィルタリング手法の一つであり、単一化に失敗する可能性の高いパス (=素性ラベル列) を事前に求めておき、単一化実行時には、それらのパスの値の単一化可能性を最初に調べることにより不要な単一化の実行を避けることができる。

**［大句構造の抑制］**大句構造の抑制は、文頭と文末に隣接しない長さ 20 単語以上の句構造を生成しない、というヒューリスティックスである。つまり、文中の大きな句構造は最終的な構文解析結果に貢献することが稀であるという経験則を応用した手法である。

**［ハイブリッド構文解析］**ハイブリッド構文解析により深い構文解析を高速化する手法が近年報告されている。前処理として、再現率は低いが適合率の高い高速な浅い構文解析を行い、深い構文解析の探索空間を大きく絞り込む。解析速度と解析精度の間にはトレードオフがあるため Penn Treebank のような実世界の大規模テキストに対してハイブリッド構文解析の効用を評価する必要がある。我々は、④で説明する非常に高速で適合率の高い CFG チャンクパーザーを HPSG 構文解析の前処理として用いた。CFG チャンクパーザーにより出力された句構造を句境界として用い、その句境界をまたがる句構造の生成を禁止することにより探索空間を大幅に絞り込む。

**ビームサーチによる高速化**

**［ビタビアルゴリズム］**解析途中の句構造は、一般に、<句構造の左端の位置, 句構造の右端の位置, 句構造を表す素性構造>で表現される。HPSG のための CYK 構文解析アルゴリズムにおいては、等価な句構造の集合を一つの句構造に縮退することにより、計算量の爆発を防いでいる。この等価な句構造の縮退の際に、句構造集合がもつ確率値のうち最も高い確率値を縮退される句構造に付与することにより確率付き HPSG 構文解析のためのビタビアルゴリズムが実現する。ビタビアルゴリズムでは最適解が保証されている。

**［ビームサーチ］**ビームサーチは最適解が保証されないかわりに大幅な高速化を実現する手法である。同じ左端の位置、右端の位置の句構造集合の中で、1) もっとも確率値の高い句構造の確率値からビーム幅 $w$ を引いた値以上の確率値を持つ句構造、および 2) 確率値が高い $n$ 個の句構造のみを解析の対象とする。近似計算により求めた外側確率を用いて句構造全体の中でもっとも高い確率値からビーム幅 $w'$ を引いた値以上の確

率値を持つ句構造のみ解析の対象とする global thresholding も実装し評価した。

**[イテレィティブ構文解析]** ビームサーチアルゴリズムでは過剰に枝刈りが行われ解が生成されないことがあるため再現率が低くなる。我々は、解析結果が出力されるまでビーム幅を広げながら繰り返し構文解析を行う手法を提案、実装した。CFG 構文解析の場合は、計算結果を記憶しておくコストと再計算コストがほぼ等価であるため、繰り返し計算を行う場合は以前に計算した結果を全て消去するほうがよいが、HPSG 構文解析の場合は、単一化による再計算はコストが高いため、以前計算した結果を再利用する。

## 評価

実験は①で説明されたように Penn Treebank の Section 02-22 を用いて生成された HPSG 文法を用い、Penn Treebank の Section 23,24 を評価対象とした。2.4 GHz の AMD Opteron CPU を搭載するサーバーを使用した。まず、上述の全ての手法を組み合わせた総合性能を評価した。表 3.1.3 は Section23 の 40 単語以下の文長に対する依存構造の適合率、再現率、F スコア(=適合率と再現率の調和平均)と一文あたりの平均解析時間を表している。

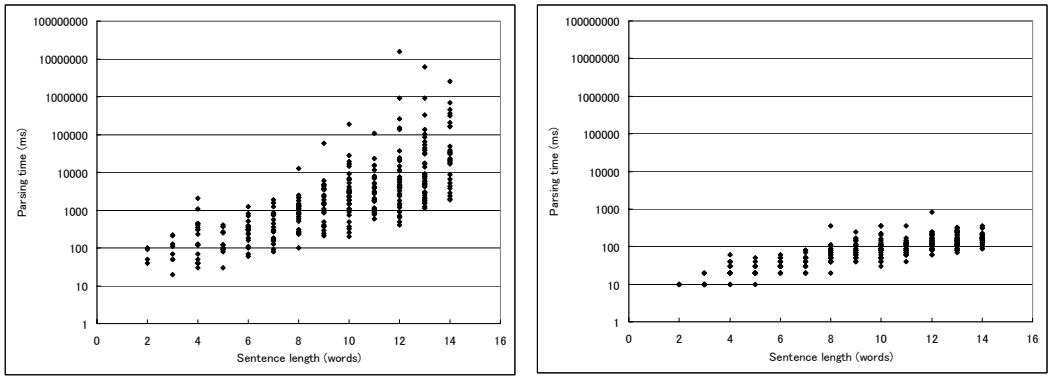
| | 適合率 | 再現率 | F スコア | 解析時間 |
|---|---|---|---|---|
| イテレィティブ | 87.6% | 86.9% | 87.4% | 360 ミリ秒/文 |

[表 3.1.3] Section23(<40 単語)に対する精度と解析速度

次に、ビタビ、ビームサーチ、イテレィティブ構文解析の性能を比較した。表 3.1.4 は、Section24 の 15 単語以下の文長に対する依存構造の適合率、再現率、F スコアと一文あたりの解析時間をビタビ、ビームサーチ、イテレィティブ構文解析の比較を表しており、図 3.1.6 は文長に対する解析時間の片対数グラフである。

| | 適合率 | 再現率 | F スコア | 解析時間 |
|---|---|---|---|---|
| ビタビ | 88.2% | 87.9% | 88.1% | 103,923 ミリ秒/文 |
| ビームサーチ | 89.0% | 82.4% | 85.5% | 88 ミリ秒/文 |
| イテレィティブ | 87.6% | 87.2% | 87.4% | 99 ミリ秒/文 |

[表 3.1.4] Section24(<15 単語)に対する精度と解析速度



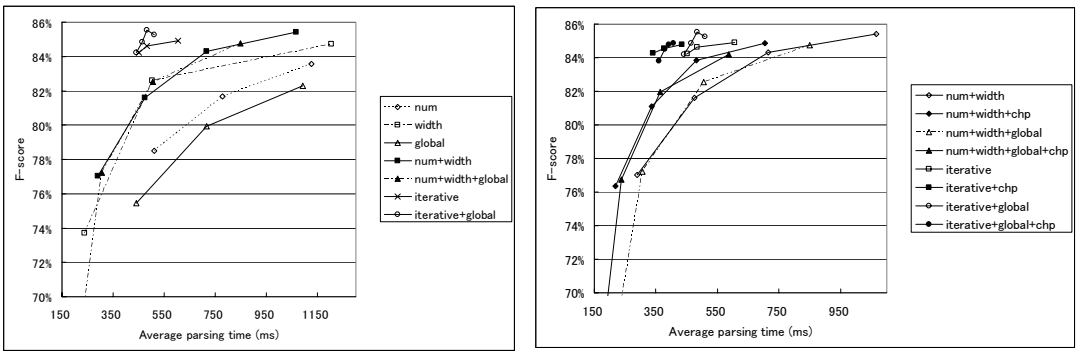

[図 3.1.6] Section24(<15 単語)に対する
ビタビの解析時間(左)と Iterative の解析時間(右)

この表から、我々の提案する HPSG 構文解析でのビームサーチの適用手法は、最適解を出力するビタビアルゴリズムと比較して、F スコアで 0.7%低くなるだけで、およそ 1000 倍の高速化を実現していることがわかる。また、同様に我々の提案するイテレィティブ構文解析により再現率がおよそ 5%ほど上昇し、F スコアでおよそ 2%ほど上昇することが確認できた。

最後に上述のそれぞれの手法の効果を Section24 の 40 単語以下の文章に対して評価した。表 3.1.5 と図 3.1.7 はその結果を表している。表は深い構文解析のための高速化手法を用いた場合(全ての高速化手法)と用いなかった場合(ビームサーチのみ)の結果を表しており、一文あたりおよそ 2.5 倍の高速化を達成していることがわかる。図の左側は、ビームサーチによる効果を表していて、ビーム幅を確率値の幅で調整したもの(width)、ビーム幅を句構造の数で調整したもの(num)、それらの組み合わせ(num+width)、global thresholding (global)、イテレィティブ構文解析(iterative)、global thresholding とイテレィティブ構文解析の組み合わせ(iterative+global)を用いたそれぞれの場合の解析時間と F スコアの関係を表している。各手法を組み合わせれば組み合わせるほど良い効果、すなわち F スコアを落とすことなく解析時間が短くなっていることがわかる。図の右側はチャンクパーザー(chp)と組み合わせたハイブリッド構文解析の効果を表しており、全ての手法に対して、ハイブリッド化による性能向上を確認できた。

| | 適合率 | 再現率 | F スコア | 解析時間 |
|---|---|---|---|---|
| 全ての高速化手法 | 85.5% | 84.2% | 84.8% | 407 ミリ秒/文 |
| ビームサーチのみ | 85.3% | 84.7% | 85.0% | 1033 ミリ秒/文 |

［表 3.1.5］各手法の効果




［図 3.1.7］Section24(<40 単語)に対する解析時間と F スコア

## ［関連発表文献］

(Ninomiya et al. 2005) Ninomiya, Takashi, Yoshimasa Tsuruoka, Yusuke Miyao and Jun'ichi Tsujii. (2005). **Efficacy of Beam Thresholding, Unification Filtering and Hybrid Parsing in Probabilistic HPSG Parsing.** In the Proc. of IWPT 2005.

(Tsuruoka and Tsujii 2005a) Tsuruoka, Yoshimasa and Jun'ichi Tsujii. (2005). **Iterative CKY Parsing for Probabilistic Context-Free Grammars.** In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2004. LNAI 3248. pp. 52-60. Springer-Verlag.

(Tsuruoka and Tsujii 2005b) Tsuruoka, Yoshimasa and Jun'ichi Tsujii. (2005). **Chunk Parsing Revisited.** In the Proceedings of the 9th International Workshop on Parsing Technologies (IWPT 2005).

### (2) 研究の成果の今後期待される効果

言語学理論に整合的な文法を現実的な応用に適用する研究は、研究開始時の５年前にはほとんどなく、ごく一部の研究グループ（ケンブリッジ大学、ペンシルベニア大学、エディンバラ大学、ザール大学、Ｐａｒｃ）が我々と同時期に研究を開始したところであった。過去の５年間の着実な研究成果の結果、ようやく、より多くの研究グループがこの方向での研究に興味を示し始めている。

上に示した我々の研究成果は、HPSG を基礎とした成果であるが、LTAG のペンシルベニア大学、ＣＣＧのエディンバラ大学、ＬＦＧの Parc においても、同様な成果を挙げてきている。意味構造レベルでの解析の精度（我々の場合には、87%）は、各グループとも同程度となっているが、新たな分野への Tuning 手法、高速なアルゴリズムの研究では、我々の研究が一歩先んじている。とくに、この２つの技術が実応用へ向けての鍵となっており、総合的な統合実験（３－５参照）でも１４億語のテキストがこの手法で解析可能であることが示せたことから、より多くの研究グループが実用的な応用に深い意味処理を使うことになると予想している。

本プロジェクトでは、生命科学の論文や新聞を対象に実用化の可能性を示したが、Enju は Generic な文法モデルとなっており、さまざまな分野に適用可能である。今後、社会科学、経済学など、幅広い分野のテキスト処理を行う基盤技術となる。

ただ、音声処理やメールの文処理に適用するには、非文法的な文の処理など、より高い耐性が必要となること、このためには、意味・知識・文脈などをより積極的に取り組む研究が必要となる。言語学的な理論に整合的で、かつ、高い耐性を持つ Enju は、これらの、より高度な研究を行うための基盤ともなる。

## ３．２　知的検索のための索引構造：領域代数と素性構造

### ［東京大学　言語処理グループ、および、オントロジーグループ］

### (1) 研究実施内容、および、成果

膨大なテキスト集合からの知的検索には、検索時点での処理と事前にテキスト集合全体に対して行っておく処理とをうまく分離する必要がある。とくに、キーワード中心の検索から一歩進んで、情報内容にまで立ち入った検索を行うには、テキストを事前に処理し、それを索引構造に反映することで、関連するテキスト集合の絞込みを行うこと、また、事前の処理結果（３－１や３－３の処理結果）を記憶しておくことで、検索時処理を軽減することが不可欠となる。事前の処理結果は、ＸＭＬを使って構造化されることから、XML タグを付与・編集するシステム、XML テキストに対する検索システム、および、素性構造データベースによる知的テキストアーカイブの技術が必要となる。

### ① XML アノテーションツール

コーパスのアノテーションは非常に多大な人的労力を要するため、エディタやワークフロー管理システムなどアノテーションを容易にするためのツールを開発することは非常に重要である。特に近年、アノテーション付きのコーパスを用いた自然言語処理が大きな発展を遂げたことを考慮すると、質の高い人手によるコーパスの開発は最も重要なタスクであるといえる。しかし、アノテーションツールの開発の問題の一つに、ツールそのもののメンテナンスコストが高い、ということがある。GUI や特殊なプログラミング言語、独自の API を用いると、ツールの開発者以外にそのプログラムの改良・保守をすることが非常に困難である。また、メンテナンスコストだけではなく、GUI を用いるプログラムは開発コストも高く、単純なブラウザを開発するだけでも数ヶ月を要し、アノテーションの書式を変更し、拡張するたびに大きな開発コストを必要とする。

我々は、XML で記述されたコーパスのメンテナンスツールを開発した。コーパスが XML で記述されていると、すでに存在する XML を管理・処理するツールを利用できること、また、それらのツールを再利用して、新しいツールを構築することが可能であり、上記のメンテナンスコストを解消するだけではなく、開発コストの面からも非常に有利である。例えば、多くのウェブブラウザは XML をサポートしているため、ヴュースタイルの設定や変更はスタイルシートとして提供が可能であり、これらのブラウザを一から作り始める必要はない。また、他のグループにより開発された XML のための商用・フリーソフトも多数存在しているためこれらのツールを利用することも可能となる。

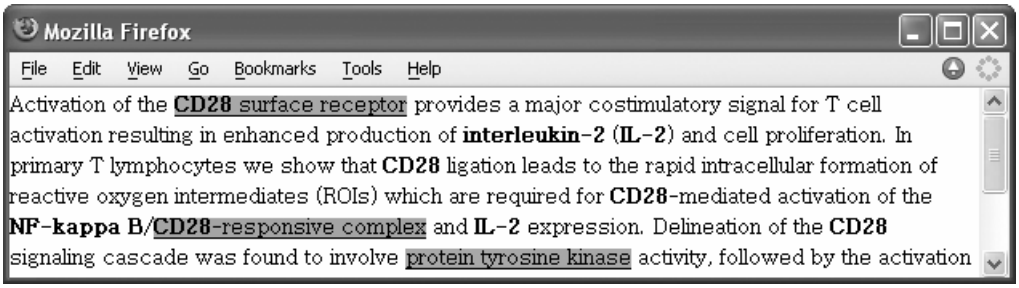
図 3.2.1 はテキストに対し蛋白質タグをアノテートしたものであり、図 3.2.2 はそれを閲覧するための CSS (Cascading Style Sheet)である。これをウェブブラウザに与えるだけで図 3.2.3 のようにウェブブラウザで閲覧することが可能となる

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/css" href="protein.css"?>
<!DOCTYPE set PUBLIC "-//TMBOOK//DTD Protein Annotation 0.1//EN" "protein.dtd">

<set>
<sentence>Activation of the <protein lex="CD28_surface_receptor" sem="family"><protein
lex="CD28" sem="molecule">CD28</protein> surface receptor</protein> provides a
major costimulatory signal for T cell activation resulting in enhanced production of
<protein lex="interleukin-2" sem="molecule">interleukin-2</protein> (<protein lex="IL-2"
sem="molecule">IL-2</protein>) and cell proliferation.</sentence>
...
</set>
```

［図 3.2.1］蛋白質のアノテーションの例

```
/* A Simple CSS for Protein Annotation */

sentence                    {font-size: 10pt;}
protein[sem="molecule"]     {font-weight: bold;}
protein[sem="family"]       {text-decoration:underline; background-color: cyan;}
protein[sem="complex"]      {text-decoration:underline; background-color: coral;}
```

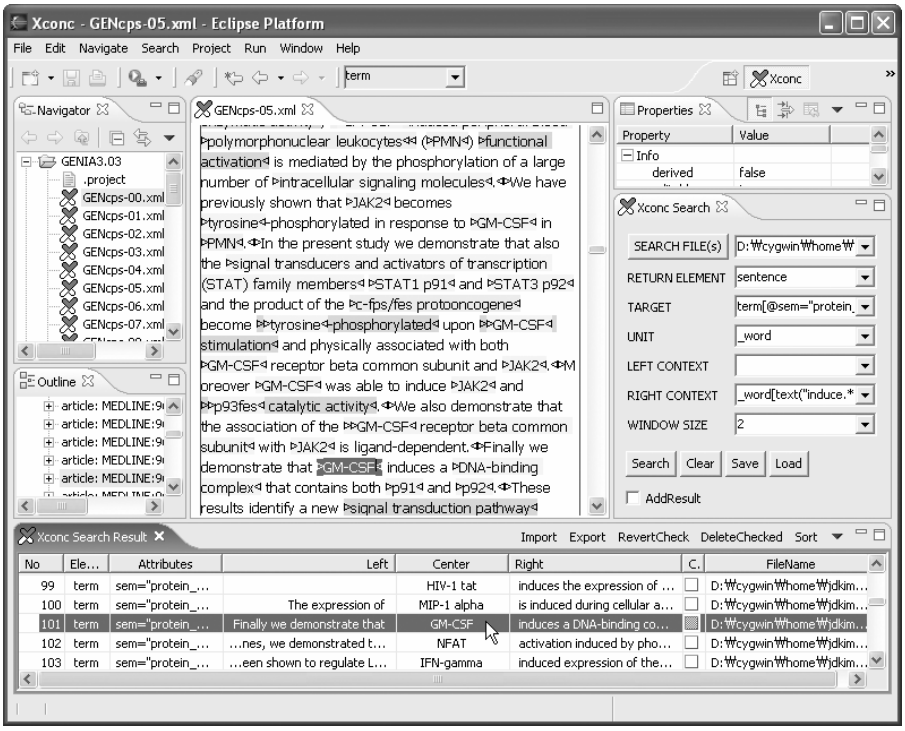
［図 3.2.2］蛋白質アノテーションのための単純な CSS




[図 3.2.3] ウェブブラウザでの閲覧

我々はこのような XML のための汎用ツールを利用する方針をとる。これらの開発は *Eclipse*[20] 上で行われた。Eclipse は広く大学、研究機関、産業界で用いられているオープンアーキテクチャーを採用しており、多数のプラグインが存在している。ここでは *VeX* XML editor[21] がコーパスファイルを編集するために使われた。DTD を与えることによって、DTD で定義された XML タグの選択やそれに付与された値の修正が可能となる。

我々は XML タグに対する検索モジュール *XConc* を開発し、Vex editor と Xconc の統合エディタを開発した (Kim and Tsujii 2005)。図 3.2.4 は統合エディタの使用画面であり、TARGET=cons[@sem=”protein_molecule”], RIGHT=_word[text(“activate.*”)], UNIT=_word, WINDOW=2 で指定されるクエリーの検索結果を示している。このクエリーにより、cons 要素の sem 属性の値が“protein_molecule”であり、その右側に“activat.*”があるテキスト領域を検索することができる。パターンの指定は、基本的には XPath と同じシンタックスを採用している。XConc と Vex editor が連動することにより検索結果に対する修正などが行える。




[図 3.2.4] Eclipse 上の Vex editor + XConc

コーパスアノテーションは長期間にわたる複数のアノテーターによる共同作業であるため、変更の履歴と共同作業を管理するヴァージョン管理システムが必要となる。CVS (Concurrent Versions System) はそのような目的に適したシステムであり、Eclipse は CVS の機能を完全にサポートしているため、統合エディタにもその機能が同時についてくる。このように汎用の XML フォーマットを用いることで、エディタ、検索システム、ヴァージョン管理システムなど莫大な開発コストを要するシステムを短期間かつ低いコストで開発することが可能であることを示した。

**［関連発表文献］**

(Kim and Tsujii 2005) Kim, Jin-Dong and Jun'ichi Tsujii. (2005) **Corpora and their Annotation**, in Text Mining for Biology and Biomedicine, to appear.

### ② 領域代数による検索システム

文書検索によく用いられるブーリアン検索では、単語の出現関係を AND、OR、NOT、前後関係で指定することができるが、XML に代表される半構造化文書に対して、構造の関係を指定することできない。領域代数は、ブーリアン検索の機能に加え、半構造化テキストに対する構造の指定を可能とする検索のための代数である。

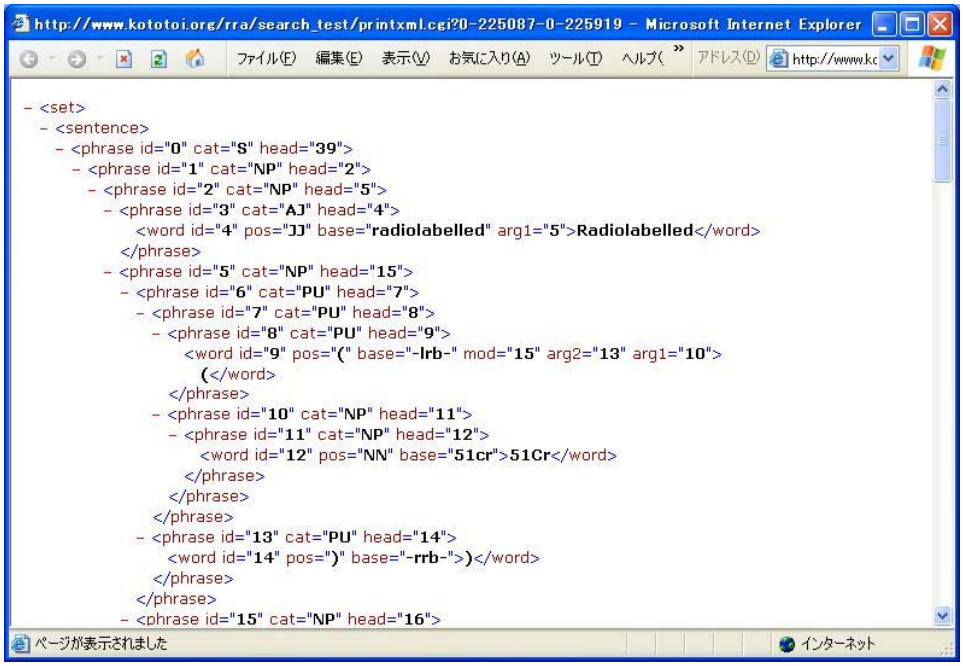
既存の領域代数では構造間の関係を指定したクエリーをシステムに与えた場合、クエリーと完全にマッチする文書領域のみ返すため、ほとんど解が得られないか、もしくは順序付けされない大量の解が得られるかのどちらかになることが多く、柔軟な検索を行うことが難しい。我々は、クエリーを部分クエリーに分解し、各部分クエリーに対し TFIDF スコアを与えることにより、ランク付き XML を実現した (Masuda et al. 2003a, Masuda 2003b)。部分クエリーに分解されているため、クエリーに完全にマッチしない場合でも解が得られる。また、部分クエリーに対して TFIDF による重み付けがなされているため、XML の部分構造に対しても重み付けがされていることになる。実験により、キーワードのみの検索や完全マッチによる領域代数の検索より、ランク付きの部分マッチによる領域代数のほうがより良い検索ができることを確認した。

本プロジェクトでは、①で説明された HPSG 構文解析器 Enju により解析された統語構造および依存構造を検索対象としている。図 3.2.5 はその XML 形式の出力の例である。XML により統語構造が表現されているが、依存構造と統語構造のタグは、タグに囲まれた領域が交差することがあるため、同一の XML で表現することができない。

図に示されているように、ここでは依存構造は ID により表現されている。このようなデータに対して、既存の領域代数を用いて統語構造と依存構造を指定する検索を行うことはできない。我々は領域代数を以下のように拡張し、図で示される XML 文書に対し、検索ができるように領域代数およびそれに対するアルゴリズムの改良を行った。

1. 既存の領域代数に対する検索アルゴリズムを改良し、既存のアルゴリズムでは正しく検索できなかった同一タグによるネスティングが存在する場合(例：phrase タグで囲まれた領域の中にさらに phrase タグで囲まれた領域が存在する)においても高速に検索を行えるようにした。

2. 領域代数の変数により、「同一の構造である」ことを示している XM に対して、それらを指定して高速に検索を行うアルゴリズムを開発した。これにより、「この動詞の主語がこのフレーズである」といったテキスト上では離れている個所にある関係を高速に検索できるようにした。




［図 3.2.5］構文解析結果が付与された Medline アブストラクト例

統語・依存構造解析結果に対して領域代数による検索を行うことによって、例えば、動詞が activate の原型、主語が proteinA、 目的語が proteinB である文書を探す、という検索要求を図 3.2.6.中のクエリーにより表現することができる。図中では、(1)で原形が activate の単語、(2)、(3)はそれぞれ proteinA、proteinB を含む phrase を示す。さらに変数$1、$2 により proteinA を含むフレーズが activate の主語(arg1)、proteinB を含むフレーズが activate の目的語(arg2)であることを表している。クエリ全体ではそれら三つの領域をすべて含む文を表している。

クエリー

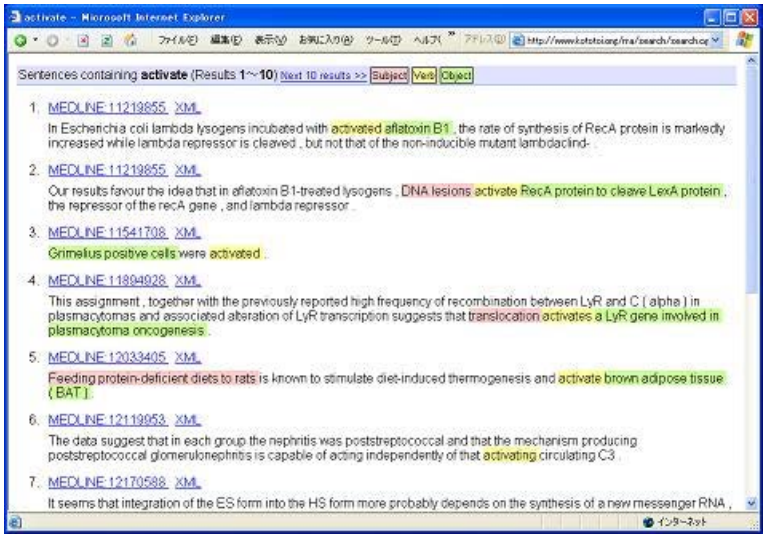
| | |
|---|---|
| [sentence] > ([word base="activate" arg1="$1" arg2="$2"] | (1) |
| & ([phrase id="$1"] > proteinA) | (2) |
| & ([phrase id="$2"] > proteinB) | (3) |

クエリー中のオペレーターの意味

| | |
|---|---|
| A > B | B を含む A の領域を返す。 |
| A & B | A で始まり B で終わる、もしくはその逆の領域を返す。意味としてはいわゆる and を表す。 |
| $1,$2 | 変数 |
| [tagname] | tagname のタグで囲まれた領域を返す。アトリビュートがある場合はその条件も満たす領域。 |

［図 3.2.6］変数付き領域代数のクエリーの例

図 3.2.7 に示される検索結果が上記のクエリーに対して得られる。図中のピンク色の領域は(2)にマッチした箇所であり、緑色の領域は(3)にマッチした箇所を示し、黄色の領域は“activate”とマッチする箇所を表している。




［図 3.2.7］クエリ“動詞:activate”に対する結果

**［関連発表文献］**

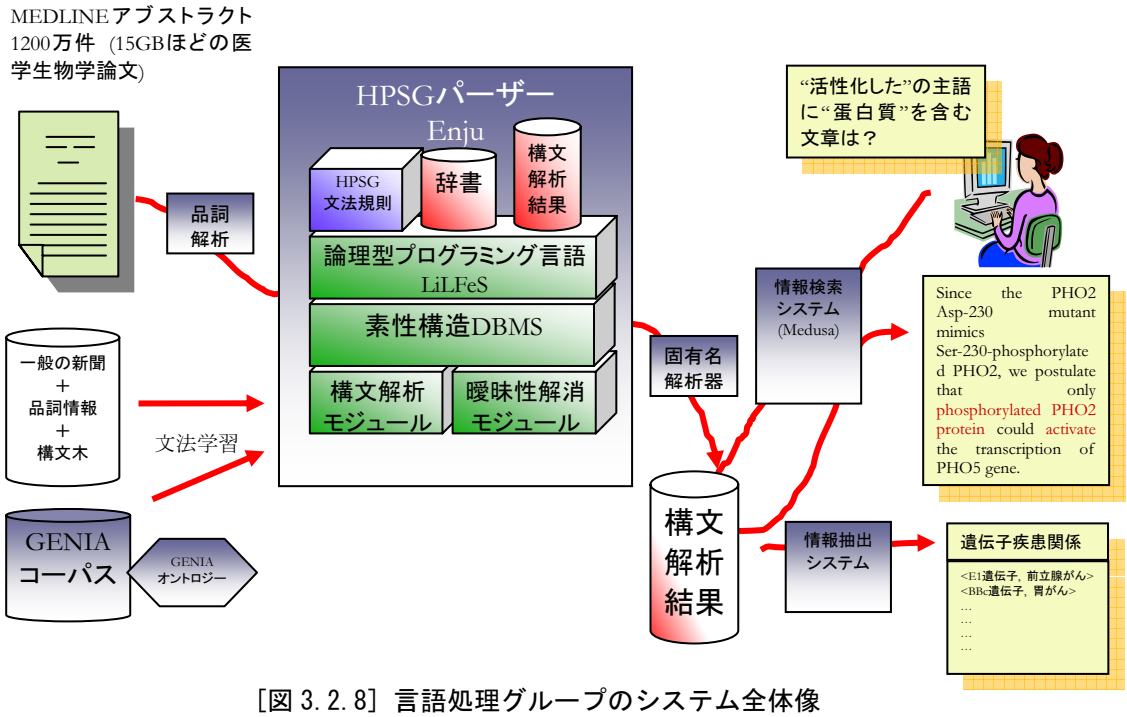
(Masuda et al. 2003a) Masuda, Katsuya, Takashi Ninomiya, Yusuke Miyao, Tomoko Ohta and Jun'ichi Tsujii. (2003). **A Robust Retrieval Engine for Proximal and Structural Search.** In the Proceedings of HLT-NAACL 2003 Short papers. pp. 58--60.

(Masuda et al. 2003b) Masuda, Katsuya. (2003). **A Ranking Model of Proximal and Structural Text Retrieval Based on Region Algebra.** In the Proceedings of the ACL 2003 Student Research Workshop. pp. 50--57.

### ③ 素性構造データベースとテキストアーカイブ

図 3.2.8 は言語処理グループのシステム全体像を表している。この中で構文解析のための中心的な役割を果たすHPSGパーザーEnjuは2つの基本システムから成っており、一つは JSPS 未来開拓プロジェクト(リーダー: 辻井潤一、1996-2000 年度)で開発された論理型プログラミング言語 LiLFeS とそのランタイムエンジン、もう一つは本プロジェクトで開発された素性構造データベースシステムである。

我々が採用した HPSG は素性構造に対する制約で定義されており、実際の文法適用は素性構造に対する単一化操作によりなされる。LiLFeS は型付素性構造を基本データとする論理型プログラミング言語であり、HPSG を記述し、素性構造を操作する構文解析器、文法学習器などのシステムの構築を容易にする。しかし、論理型プログラミング言語 LiLFeS は、プログラムで記述されたデータやメモリ上の一時的データを操作することは容易にできるが、永続的にファイルシステム上に蓄積されるデータベースを管理する機能は実現されておらず、自動的に学習された辞書項目を格納したり、素性構造で表現される構文解析結果を格納することは困難であった。



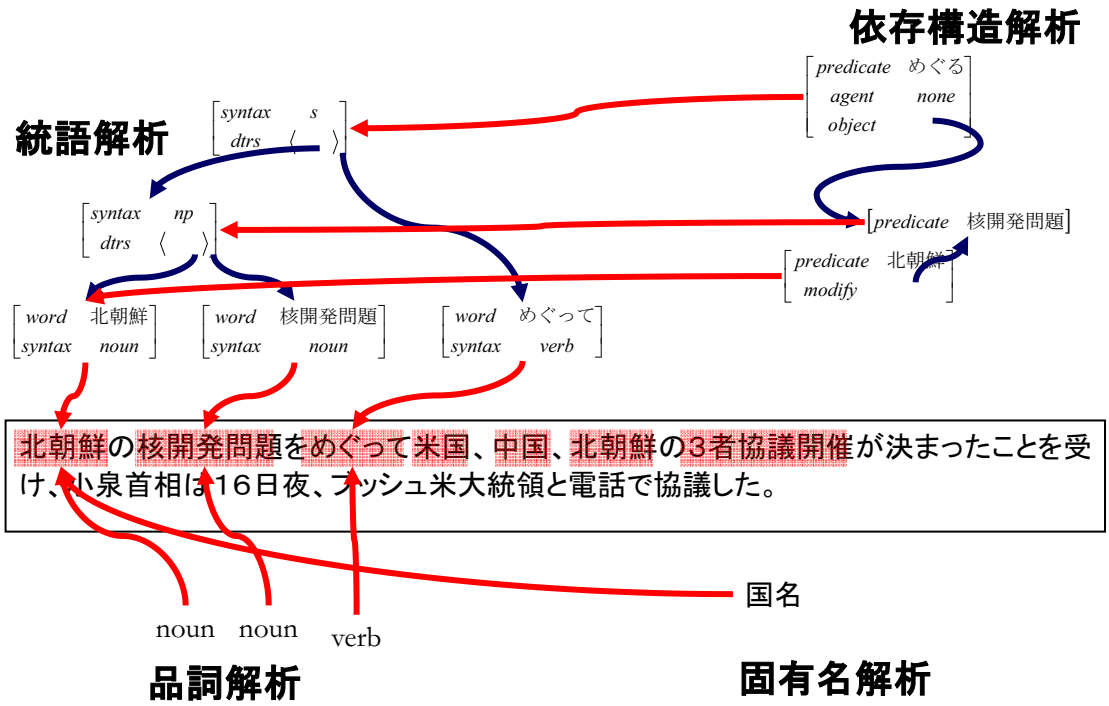

［図 3.2.8］言語処理グループのシステム全体像

我々は、まず、C++から操作できる Persistent Standard Template Library (PSTL)を開発し、それを用いて、LiLFeS と C++で操作可能な素性構造データベース管理システムを開発した。我々は実用的なシステムの開発を目指しており、プロトタイプシステムは LiLFeS で記述されることが多いが、実用的なシステムは主に C++で再実装され、文法適用などの素性構造操作が必要な場合には C++から LiLFeS ランタイムモジュールが呼び出される。そのため C++と LiLFeS の両言語から容易に操作できるデータベース管理システムが望ましい。我々はまず C++で広く使われている Standard Template Library (STL) のファイルシステム版を作成した。STL はリストやベクター、ハッシュ、ツリーなどの一般的なコンテナタイプのデータ構造を C++上で容易に取り扱うことを可能とするテンプレートライブラリであり、C++の標準ライブラリとして、多くの C++コンパイラに用意されており、また、非常に広く使われている。しかしながら、ファイルシステム上に構築される STL で本格的な使用に耐えうるスケーラブルな実装はほとんどなく、我々はファイル上に永続的にデータを格納できる STL (PSTL) を設計、開発した。LiLFeS 上のデータベース管理システムは PSTL を用いて実装されており、素性構造を格納するベクター、リスト、および、等価な型付素性構造をキーとして型付素性構造を検索できるハッシュとツリーが実装されている。また、単一化可能な型付素性構造をキーとして検索する索引機能 (Ninomiya et al. 2002)も研究、開発し、実装されている。

索引機能により、素性構造データベースから、単一化可能な型付素性構造を高速に検索することが可能であり、実験により、単純な検索に比較し数十倍の速度で検索できることが確認された。素性構造はラベル付、根付グラフ構造であり、根からあるノードまでの枝に付与されたラベルの列をパスと呼んでいる。索引機能を用いた検索は、「単一化可能な素性構造は全てのパスに対してお互い単一化可能な値が付与されていなけれ

ばならない」という単一化のための必要条件に基づいている。素性構造がクエリーとして与えられたあと、まず、もっとも解の候補を大きく絞ることができるパスを動的に数個決定し、次にそれらのパスの値から実際に解の候補を大きく絞り込む。解を単一化により求めることにより、部分的知識を記述したクエリーを記述でき、検索順序に依存しない推論が可能となり、知的検索システムとしての良い性質を備えているといえる。

HPSG の構文解析結果は素性構造で表現される統語・依存構造であるため、それらの解析結果を直接素性構造データベースに格納し、検索することができる。我々は素性構造をテキストに対し付与するシステムを考案、開発した (Ninomiya et al. 2005)。図 3.2.9 はその概念図を表している。統語解析は構文木を表しており、葉の部分には単語の領域が対応している。依存構造は統語解析の句構造および単語に対応しており、句構造、単語間の関係を示している。また、品詞タガーや固有名解析タガーの結果も同様にテキストに対して付与することができるため、上述の領域代数の機能と組み合わせることにより、様々な解析器の結果を利用した検索が可能となる。統合システムでは、このシステムを用いて、Medline アブストラクト全体(約１５００万抄録、約１４億語)を解析した結果を素性構造データベースに格納しており、指定された意味構造に対し、1 秒以内に検索結果が得られることを確認した。




［図 3.2.9］素性構造付きテキストアーカイブの例

**［関連発表文献］**

(Ninomiya et al. 2002) Ninomiya, Takashi, Takaki Makino and Jun'ichi Tsujii. (2002). **An Indexing Scheme for Typed Feature Structures.** In the Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). pp. 1248-1252.

(Ninomiya et al. 2005) Ninomiya, Takashi, Yusuke Miyao and Jun'ichi Tsujii. (2005). **A Persistent Feature-Object Database for Intelligent Text Archive Systems.** In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2004. LNAI3248. pp. 197--205. Springer-Verlag.

### (2) 研究成果の今後期待される効果

テキストにＸＭＬを使ってアノテーションを行い、それを中核にして情報内容の統合を行う試みは、ＩＢＭのＵＩＭＡ(Unstructured Information Management Architecture)の構想など、近年盛んに提案されている。我々のここでの研究は、それら一連の研究に領域代数、素性構造の理論を導入することで、理論的にしっかりした土台を与えようとしたものである。我々の作成したシステムは、ＩＢＭのＵＩＭＡが、Semantic Retreival System としてシステム外においている コンポーネントと対応しており、彼らが開発している検索系よりも豊富な機能を提供している。

この３つの研究成果は、いずれも、言語リソースの作成（３－４節)、統合検索システム（３－５節）のために実際の場面で、しかも、大量のデータを対象とした場合での有効性が確認されている。今後、ほかの分野での同種の試みを行う有効なツールキットとなることが期待できる。とくに、領域代数・素性構造に基づく検索系は、ＵＣＳＤ(米国)とリバプール大学（英国）と共同で、かれらのテキストベースシステム（Cheshire)との性能比較を行うことを計画している。Cheshire は、米国・英国のディジタルライブラリー研究でも広く使われおり、広い研究集団に影響を与えることが期待できる。

## ３．３　テキスト処理のための機械学習手法と適用アルゴリズム

［東京大学　言語処理グループ］

### (1) 研究の実施内容と成果

近年、機械学習は、テキスト処理だけでなく、データマイニング・音声認識・ロボット研究など、非常に幅の広い分野に適用され、大きな成果を挙げている。ただ、言語処理・テキスト処理に適用するためには、これらの分野の特質にあった手法を開発する必要がある。とくに、（１）大量の非構造化データの扱いが高効率な手法（学習に要する時間、および、処理速度)、（２）単語 n-字組のような大量の素性で、かつ、スパースなデータに対して Robust な特性を示す手法、（３）言語の特徴である一次元の系列データに向いた手法、（４）言語解析のアルゴリズムと機械学習手法とを有機的に組み合わせる手法、の開発が不可欠となる。

本研究プロジェクトでは、上記（２）のために非等号制約を認める高効率なＭＥ(Maximum Entropy)モデル Kazama 2005)、および、（４）として、Enju の確率モデルを開発した（３－１節参照)。このＭＥモデルを用いて、かつ、CRF と同様な大域的な最適化を達成する高効率な手法(Easiest-First)を開発した。

本節では、上記（３)・（４）のために開発した手法を中心に報告する。

#### ① 系列タグ付け問題とその効率的な処理アルゴリズム

自然言語処理の基盤的技術である、品詞タグ付け、固有表現認識、shallow parsing な

どの処理は、多くの場合、単語列（文）に対するタグ付け処理という形で定式化することができる。そのような問題に対するアプローチとして近年注目を集めているのは、文全体を直接モデリングの対象とし、log-linear モデルや max-margin などの手法によってパラメータを最適化するというアプローチである。近年、固有表現認識などでよく使われる Conditional Random Field (CRF) は、このアプローチに属する。

CRF や max-margin 法では、タグ列に関する特徴量のタイプを制限しないでモデル化しようとすると、学習に必要な計算コストが爆発してしまうため、特徴量の形を制限し、ダイナミックプログラミングが可能な形にすることで、多項式時間で学習・タグ付けができるようにする。それでも、たとえばタグ列に関する 2 次の特徴量を利用しようとすると、計算コストが大きすぎて CRF で実用的な品詞タガーを構築することは難しい。Shallow parsing においても、対象を名詞句のみに限定することで、やっと学習が可能というのが現状である。

本研究グループでは、シーケンスに関する高次の特徴量が利用でき、しかも、高効率で学習とタグ付けが可能な手法を開発した (Tsuruoka and Tsujii 2005a)。

### Bidirectional Inference

単語列に対するタグ付けの問題は、$P(t_1...t_n \mid o)$を最大化する問題だと考えることができる。ただし$t_1...t_n$はタグ列、$o$は単語列とそれに関する特徴を表す。これは、次のように分解することができる。

$$P(t_1...t_n \mid o) = \prod_{i=1}^{n} p(t_{i:} \mid t_1...t_{i-1} o)$$

ここで、タグに関して１次マルコフモデルを仮定する、つまり、あるタグから２つ以上離れているタグは、そのタグの確率分布に影響しないと仮定すると、次の式が得られる。

$$P(t_1...t_n \mid o) = \prod_{i=1}^{n} p(t_{i:} \mid t_{i-1} o)$$

これが MEMM でよく利用される、左から右への分解である。もちろん、右から左に分解することもできる。このように分解することで、一つ一つのタグを予測することが単純な分類問題になり、既存の機械学習手法をそのまま利用することができる。

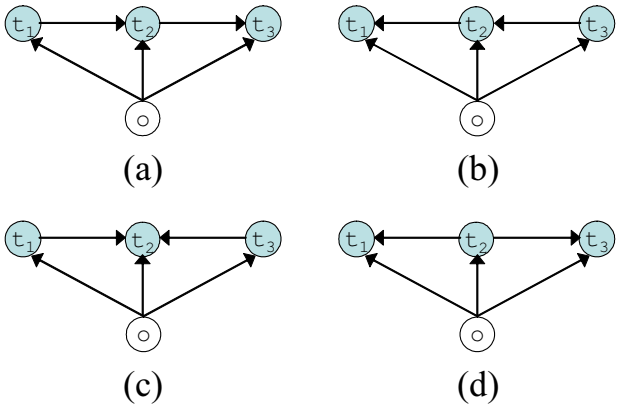
重要な点は、左から右への分解をした場合、機械学習の特徴量としては、予測対象の左側のタグ情報しか使えないことである。逆に、右から左への分解にした場合は、予測対象に右側のタグ情報しか利用することができない。

分類をする機械学習の立場から言えば、特徴量としてはなるべく多くの情報を利用したい。特に、近接するタグに関する情報は非常に重要である。このような動機から、Toutanova らは、以下のようなスコアをタグ列に対して考えることで、個々のタグの予測において両側のタグ情報が利用できるようにして、英語の品詞タグ付けで最高精度達成している。

$$score = \prod_{i=1}^{n} p(t_{i:} \mid t_{i-1} t_{+1} o)$$

ただし、この定式化にはいくつかの問題点がある。一つには、確率の条件付けに、一つのタグ情報が２重に使われているため、情報の一種の２重カウントが起きていることである。このため、タスクの性質によっては精度が大きく低下することがある。また、この式における最適タグ列を求めるためには、探索を行う必要があり、高速なタガーを実現することが難しい。

本提案手法では上記のようなスコアを考えるのではなく、すべての分解の構造を考慮することにより、両側のタグ情報を利用する。例えば、３単語からなる文の確率の分解の構造をすべて考えると図 3.3.1 のようになる。




［図 3.3.1］ タグ情報の依存関係のネットワーク

これらの構造は、以下の４つの式に対応する。

$$(a)\ P(t_1...t_3 \mid o) = P(t_1 \mid o)P(t_2 \mid t_1 o)P(t_3 \mid t_2 o)$$
$$(b)\ P(t_1...t_3 \mid o) = P(t_3 \mid o)P(t_2 \mid t_3 o)P(t_1 \mid t_2 o)$$
$$(c)\ P(t_1...t_3 \mid o) = P(t_1 \mid o)P(t_3 \mid o)P(t_2 \mid t_3 t_1 o)$$
$$(d)\ P(t_1...t_3 \mid o) = P(t_2 \mid o)P(t_1 \mid t_2 o)P(t_3 \mid t_2 o)$$

式（c）に対応する構造を選択すれば、２番目の単語のタグを判定する際に、両側のタグ情報が使えることがわかる。提案手法では、分解の構造とタグ列をすべて考え、最も確率の高いタグ列と構造を出力する。一般に、ｎ単語からなる単語列に対しては 2^(n-1)個の分解方法があるため、ナイーブなアルゴリズムでは、単語列長に対して指数オーダーの計算量が必要になることになるが、我々は、ダイナミックプログラミングを行うことで多項式時間で最適解を見つけ出すアルゴリズムを開発した。

さらに、greedy な方法として、確率値の高いタグから決定して構造を決めてゆく手法（Easiest-first）を提案している。Easiest-first 手法では、最も確率の高い構造が選ばれる保証はなくなるかわりに、きわめて高速なタグ付けが可能になる。

表 3.3.1～表 3.3.3 に提案手法の精度とタグ付けの速度を示す。提案手法によって両側のタグ情報を利用することで一方向 MEMM から大きく精度が向上していることがわかる。また、Easiest-first 手法では、ほとんど精度を落とすことなく大幅な速度向上を実現している。絶対性能としても、SVM などを利用した他のタガーと比べて同等かそれ以上の精度を、小さな計算コストで実現していることがわかる。

| 手法 | 精度（%） | 速度 (tokens/sec) |
|---|---|---|
| Left-to-right (Viterbi) | 96.92 | 844 |
| Right-to-left (Viterbi) | 96.89 | 902 |
| Dependency Networks | 97.06 | 1,446 |
| Easiest-last | 96.58 | 2,360 |
| Easiest-first | 97.13 | 2,461 |
| Full Bidirectional | 97.12 | 34 |

[表 3.3.1] 手法による精度の違い(Wall Street Journal corpus の Section 22-24)

| 手法 | 精度（%） |
|---|---|
| Dep. Networks (Toutanova et al., 2003) | 97.24 |
| Perceptron (Collins, 2002) | 97.11 |
| SVM (Gimenez and Marquez, 2003) | 97.05 |
| HMM (Brants, 2000) | 96.48 |
| Easiest-first | 97.10 |
| Full Bidirectional | 97.15 |

[表 3.3.2] 品詞タガーの精度 (Wall Street Journal corpus の Section 22-24)

| 手法 | 再現率 | 適合率 | F 値 |
|---|---|---|---|
| SVM (Kudoh and Matsumoto, 2000) | 93.51 | 93.45 | 93.48 |
| SVM voting (Kudo and Matsumoto, 2001) | 93.92 | 93.89 | 93.91 |
| Regularized Winnow (basic features)(Zhang et al., 2002) | 93.60 | 93.54 | 93.57 |
| Perceptron (Carreras and Marquez, 2003) | 93.29 | 94.19 | 93.74 |
| Easiest-first (IOB2, second-order) | 93.59 | 93.68 | 93.63 |
| Full Bidirectional (Start/End, first-order) | 93.70 | 93.65 | 93.70 |

[表 3.3.3] Chunking の精度 (Wall Street Journal corpus の Section 20)

### ② 部分タグ列分類問題としての定式化とアルゴリズム

タグ列に対する機械学習手法としては、先に述べたように、MEMM や CRF による方法が広く使われている。これは、タグ列同士の関係が、全体のタグ列を決定する上で大きな影響をもっているからだといえる。

それに対して sliding window 方式では、タグ同士の依存関係を無視し、問題を、単なる部分タグ列の分類問題としてとらえる。固有表現認識などのタスクにおいては、MEMM や CRF であれば、いわゆる BIO タグ表現を利用して、本来ひとつながりのチャンクをばらばらのタグに変換する必要があるが、sliding window 方式であれば、チャンクをそのまま機械学習の学習対象とすることができる。そうすると、チャンク全体にわ

たる特徴量などを定義することが可能になり、特にバイオ分野における NER のような、チャンクが長い場合には、有効な特徴量を多く利用することができるという利点がある。このような、部分シーケンス全体にわたる特徴量を定義できる手法としては、Semi-Markov CRF なども提案されているが、計算コストが大きいことなどから、バイオ分野の NER などで実用には使われていない。

Sliding window 方式の欠点は、基本的には、n 単語の文において n(n-1)/2 個の部分シーケンスを数え上げる必要があるために、学習のための計算コストが大きくなりすぎるという点にある。そこで我々は、Naïve Bayes 分類器をフィルターとして利用して、それを通過したものだけを学習対象とする手法を試みた。

表 3.3.4 に提案手法をバイオ分野の固有表現認識に適応した結果を示す。多くのヒューリスティクスを駆使して最高性能を達成しているタガーには劣るものの、高い性能で固有表現認識が可能であることがわかる。また、提案手法を CFG パージングに適用した結果、非常に高速な CFG パーザを構築することが可能となった（1 文平均 14 ミリ秒、F 値 85）（Tsuruoka and Tsujii 2005b)。

| | 再現率 | 適合率 | F 値 |
|---|---|---|---|
| Support Vector Machine & HMM (Zhou 2004) | 76.0 | 69.4 | 72.6 |
| 提案手法 | 72.8 | 68.8 | 70.7 |
| MEMM (Finkel et al., 2004) | 71.6 | 68.6 | 70.1 |
| CRF (Settles., 2004) | 70.3 | 69.3 | 69.8 |

**［表 3.3.4］ バイオ用固有表現認識の精度**

**［関連発表文献］**

(Tsuruoka and Tsujii, 2005a) Tsuruoka, Yoshimasa and Jun'ichi Tsujii. (2005). **Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data.** In the Proceedings of HLT/EMNLP 2005. (to appear)

(Tsuruoka and Tsujii 2005b) Tsuruoka, Yoshimasa and Jun'ichi Tsujii. (2005). **Chunk Parsing Revisited.** In the Proceedings of the 9th International Workshop on Parsing Technologies (IWPT 2005). (to appear)

### (2) 研究成果の今後期待される効果

機械学習が非常に幅の広い応用は周知のことであるが、実際の問題に適用するに当たっては、当該問題の特殊性を十分に考慮した手法の開発と、機械学習の結果を有効に活用できるアルゴリズムの確立が不可欠となる。言語処理・テキスト処理の特殊性を取り込んだ機械学習手法の研究は、本プロジェクトの理論での柱の一つであり、その成果は、非統合制約を認めるＭＥ、文解析と確率モデルとの統合という成果をあげている。

本節で述べた２つの手法は、品詞タグ付け・ＮＥＲ・ＮＰ Chanker など、言語処理の広い範囲の問題を解くためのアルゴリズムと確率モデルとの統合を行うものである。この２つの手法は、系列タグづけ問題に局所的な情報だけでなく大域的な情報を取り込むことを意図する点で、急速に広まっているＣＲＦと同じ方向性を持つが、計算的にははるかに簡便であり、実用上、非常に有効な手法となっている。実際に、我々が開発し

たGENIAタガー(生命科学分野のテキストに対する品詞タグ付け)、ＮＥＲ、表層パーサは、ＣＲＦと同等な性能をより高速に実現していることを確認している。

テキスト処理・言語処理の問題で、系列タグ付け問題として定式化できる問題は非常に多い。我々の成果は、これまで、個別の問題として、個々に効率化と高精度化を目指していた一連の問題を同一の枠組みでみること、かつ、State-of-Artsの精度をはるかに効率的に実現するもので、今後、さまざまな問題に適用されることが期待できる。

## ３．４　オントロジーと言語リソースの構築
## ［東京大学　言語処理グループ］

### (1) 研究の実施内容と成果

言語処理・テキスト処理の研究を系統的に行うためには、机上の空論的な理論ではなく、現実の言語使用の実態に即した、ボトムアップ的な方法論が不可欠となる。３－１節でのコーパスに基づく文法モデルの開発は、理論からのトップダウン的なモデル化とデータからのボトムアップ的な手法を結びつけたものであった。ただ、ボトムアップ的な方法論を有効に適用するためには、テキストをそのまま集めた「生コーパス」だけでなく、理論から派生するさまざまな情報（品詞、構文木など）を付加し、理論からの派生概念が実態のテキストにどのように現れるかを見る必要がある。

とくに、オントロジー・意味といった、直接観察できないものと言語とを結びつけるには、品詞・構文木といった言語学の理論からの派生概念だけでなく、分野の知識（オントロジー）から派生する情報も付加しなければならない。このような「タグ付きコーパス」、あるいは、「注釈つきコーパス」は、機械学習ベースの言語処理（３－３節参照）の発達に伴って、その重要性が高まっている。

本プロジェクトでは、オントロジーを使った知識処理とテキスト処理の融合を目指す立場から、オントロジー研究への関心が高く、かつ、実際の場面での知的テキスト処理への要求も高い生命科学分野に焦点をあって、テキストへのオントロジー情報を埋め込む作業を行った。その成果は、生命科学分野でのオントロジー(ＧＥＮＩＡオントロジー)とそれに基づく意味アノテーションつきコーパス（ＧＥＮＩＡコーパス）として公開されている。また、生命科学分野のテキストは、通常の言語使用とは異なる現象（多数の専門用語と略語、多様な並列句、数式・化学式の混入、など）が多くみられること、構文構造の統計的な偏りなどの理由から、品詞タグ、構文木など、言語情報の付加作業を行った。これらの結果は、３－１、３－３、３－６節の各研究課題で活用されただけでなく、世界の研究グループ（２４０超）のグループによって使われている。以下では、コーパスの概要と各レベルの注釈付加について述べる。

#### ① GENIAコーパスの概要

われわれは、機械学習ベースの自然言語処理の手法を応用するための学習および検証データとして、MEDLINEデータベース上のアブストラクトに専門用語・相互作用イベント・品詞・構文木をタグ付けしたコーパス（GENIA コーパス）を開発した。新聞な

どのテキストに対してすでに実用可能な性能を示す品詞付けについても、この分野のテキストに対しては改良の余地があることから、われわれはMEDLINEアブストラクトに対して考えられる限りの情報を陽にマークアップしていこうという方針でコーパスを作成する。同一のテキストに各種の情報をタグ付けすることによって情報相互間の関係、たとえば用語抽出のためにどんな言語知識が手がかりとなりうるか、を観察するのに役立てることも目的としている。

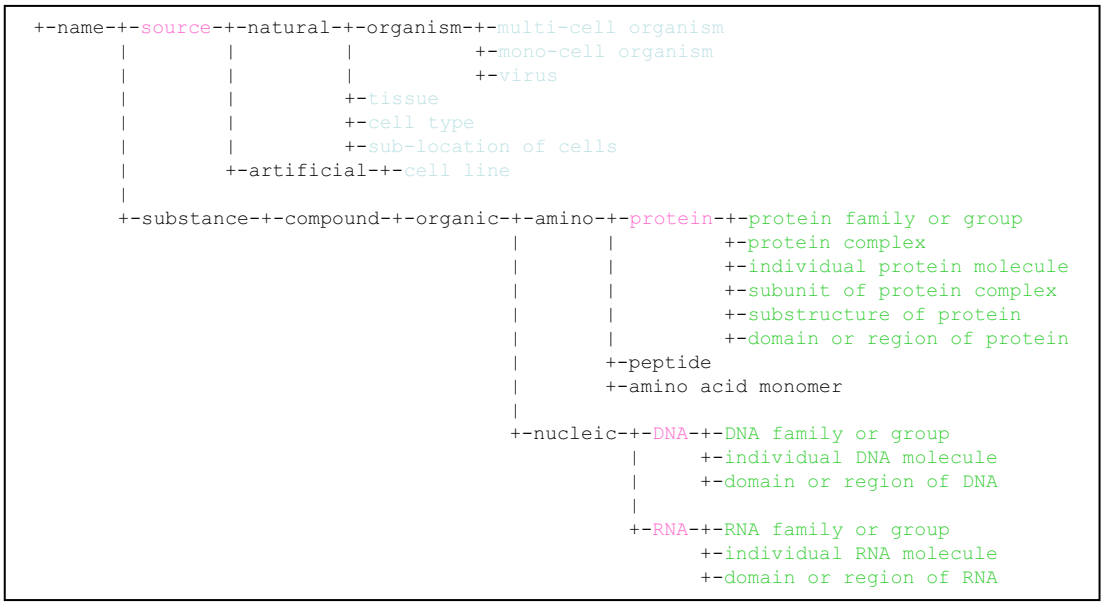
対象とするテキストは、MEDLINE データベースから human（ヒト）、 transcription factors（転写因子）、 blood cells（血球細胞）の 3 つをキーワードとして検索された結果のアブストラクト群である。これは、生命科学分野の中でもかなり狭い研究領域になっているため、辞書ベースのアプローチに利用するにはそのカバレージは十分ではない。しかし、領域を広く設定しすぎると、専門用語、特に物質名が多岐にわたり、機械学習ベースのアプローチに有効な特徴を捉えることができなくなる。一方、生命科学分野特有の言語学的な特徴は、この絞り込んだアブストラクト群からも十分に捕らえることができると考え、この領域のアブストラクト群を用いている。

現在、われわれのグループでは、専門用語(2000 アブストラクト約 40 万語)、品詞（専門用語コーパスと同一セット)、構文木（専門用語コーパスのサブセット 500 アブストラクト約 10 万語）をタグ付けしたコーパスを公開し、さらに物質の相互作用イベントをタグ付けしたコーパスを作成している。また、Institute for Infocomm Research(シンガポール)における MedCo プロジェクトでは GENIA コーパスの一部(670 アブストラクト)に照応関係（代名詞などの参照関係）をタグ付けしている。以下では個々のタグ付けの方針・タグ設計・およびタグ付けの際に生じた問題点を述べる。

### ② GENIA 専門用語コーパスとＧＥＮＩＡオントロジー

GENIA 専門用語コーパス(Kim et al, 2003)は、物質とその所在（ソース）の名の位置を同定するとともに、各々の用語についてタンパク質名・細胞名などのクラス分けをする。

このクラス分けのために、我々は GENIA オントロジーと呼ぶ小規模な分類体系(図 3.4.1)を構築した。GENIA オントロジーは Substance(物質名)、Source（所在）および Other(そのほか) というサブコンポーネントからなる概念の階層関係の木構造で、専門用語にはこの葉ノードのいずれかを意味クラスとして付与している。コーパスの設計段階では、葉ノードだけでなく、意味クラスとしていずれのノードを選択してもよく、また複数の意味クラスを付与することも考慮に入れていた。しかし実際タグ付け作業を進めてみると、作業者に十分な背景知識があれば、葉ノードを割り当てることが可能であることがわかった。さらに、専門用語には当然多義語も存在し、同一の表現で複数の意味クラスに属する可能性があるが、それぞれの出現箇所では、その文脈に依存して一つの意味クラスを割り当てることが可能であった。ただし、現在の GENIA オントロジーでは、物質の化学的な性質のみに基づいて分類しており、機能に関する分類を行っていない。今後、生体内での役割などの物質の機能に関する概念を定義し、現在タグ付けされている用語に対してさらに属性を追加していく予定である。




［図 3.4.1］ GENIA オントロジー（物質的な概念の分類）

### ③ GENIA 品詞コーパス

MEDLINE アブストラクトに対して、汎用の品詞タガーである Junk Tagger を適用すると、Wall Street Journal に対して 96.8%あった精度が 83.5%まで落ちる。そこでわれわれは専門用語コーパスと同じアブストラクトセットに対して品詞タグを付与したコーパスを作成した(Tateisi et al, 2004)。品詞コーパスではアブストラクトの各単語に Penn Treebank(ペンシルバニア大学で作成され広く使われている品詞・構文タグ付コーパス)のセットに基づく品詞を割り当てる。タグ付与は、まず前処理として Penn Treebank のトークナイザーを用いて機械的にトークンに分割、Junk Tagger で初期タグ付けを行い人手で修正する、という手順をとった。ここで、生命科学分野の論文に特有と思われる２つのの問題が生じた。

１つは、物質名に、通常のテキストならば独立のトークンとみなされるコンマや括弧などの記号がふくまれることがある(例：2,3,7,8-tetrachlorodibenzo-p-dioxin)ことであり、もう１つは、修正を行うアノテータが専門用語中の語の品詞づけに迷うことが多いことであった。特に、この分野では人名などのいわゆる固有名詞ではない物質名や細胞株名が大文字で始まる(例：物質名 Jun、細胞株名 Jurkat)特性を持ち、固有名詞と普通名詞の区別に迷うことが多かった。しかし、これらの問題点は専門用語に集中しており、ほとんどは文脈に依存せずに品詞を決めることができる。このことから、１）専門用語コーパスから専門用語と認定された部分を切り出してそれらの語を構成する各単語に人手で品詞をつけた辞書を作り、２）それをもとにトークン分割を修正するとともに専門用語を構成する各トークンにあらかじめ機械的に品詞を割り当て、３）Junk Tagger はそのタグと矛盾しないように残りの部分に品詞タグをつけ、４）人手で修正する際にはあらかじめ機械的に割り当てた品詞は正しいものとする、とすれば安定した品詞付けが行える。また、固有名詞と普通名詞の区別は構文解析など品詞タグを利用するプロセスに

はあまり重要ではなく、細かい基準を作ってこんらんするよりも安定して判断できることのほうが重要であると考え、論文の著者名などの人名、研究機関名などをのぞき普通名詞として扱うように Penn Treebank の品詞付け基準を修正することにした。
品詞コーパスは Penn Treebank に準拠した形式と、専門用語タグのついたコーパスに品詞タグの情報をマージした XML 形式の 2 つの形式で公開した。後者の形式では、専門用語は一つ以上のトークンからなるものと仮定したが、論文アブストラクトでは限られた文字数で論文の概要を説明しようとすることから、等位接続の頻度が高く、特に、専門用語についてトークンよりもさらに狭い単位で共通する接頭辞・接尾辞などを共有する語同士の等位接続がある（例：homo- or heterodimers)。このような場合、専門用語タグでトークンが分割されることが起こる。このように分割されたトークンに対しては品詞の変わりに”*”を割り当てている（図 3.4.2)。

```
<abstract>
…
<sentence><cons lex="IL-2-mediated_T_cell_proliferation" sem="G#other_name">
<cons lex="IL-2" sem="G#protein_molecule">
<w c="*">IL-2</w></cons><w c="JJ">-mediated</w>
<cons lex="T_cell" sem="G#cell_type"><w c="NN">T</w>
<w c="NN">cell</w></cons> <w c="NN">proliferation</w></cons>…
…
</abstract>
```

[図 3.4.2] GENIA 品詞コーパス

### ④ GENIA 構文木コーパス

構文木コーパス(Tateisi et al, 2005)は専門用語コーパスの対象アブストラクトのサブセットに対して、GDA-DTD を拡張した DTD を用いて PennTreebank と同様の構造を XML 形式で付与したものである（図 3.4.3)。

```
<S><PP>In <NP>the present paper </NP></PP>, <
NP-SBJ id="i55"><NP>the binding </NP><PP>of <NP>a [125I]-labeled aldosterone
derivative </NP></PP><PP>to <NP><NP>plasma membrane rich fractions </NP><PP>of
HML </PP></NP></PP></NP-SBJ><VP>was <VP>studied <NP NULL="NONE"
ref="i55"/></VP>
</VP>.</S>

(S (PP In (NP the present paper)), (NP-SBJ-55 (NP the binding) (PP of (NP a
[125I]-labeled aldosterone derivative)) (PP to (NP (NP plasma membrane rich
fractions) (PP of HML)))) (VP was (VP studied *-55)).)
```

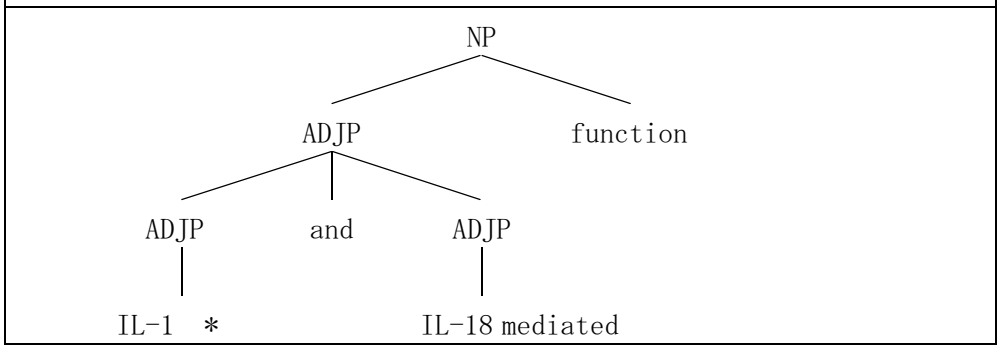
[図 3.4.3] GENIA 構文木コーパス。上は XML 形式。下は Penn Treebank 形式

タグ付けの基準はほぼ Penn Treebank に基づいているが生命科学分野では専門用語が時には前置詞句を含む場合があるなど長い（多数の単語からなる）場合が多く、専門用

語内部の構造をどこまで、あるいはどのように分析するかを決めるのは困難であると予想された。品詞付けの場合と同様に専門用語内部の構造は文脈に依存しないものと考えられる。したがって、専門用語に関しては文脈から独立に分析した方が、同じ用語に対しての分析がゆれる危険がなくなってよいと判断し、構文木タグ付けの際は専門用語内の構造は分析しないこととした。タグ付けは完全に人手で行い、アノテータにはタグ付けに迷った点についてできるだけコメントしておいてもらうようにした。作成した XML ファイルは、自動的に Penn Treebank 形式に変換し、両方の形式を公開している。

作業を通して、等位接続の構造付けが困難であることがわかった。品詞コーパスの項でも述べたように、論文アブストラクトでは一部を省略した等位接続が多い。そのような省略を復元するには文の構文のみならず意味に関する十分な理解が必要である。たとえば図 3.4.4 は同じ句“IL-1 and IL-18-mediated function”に対する２つの構造化を示す。

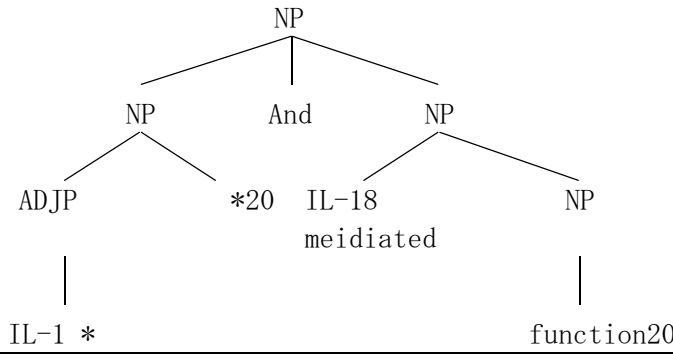
```
<NP>
<ADJP SYN="COOD">
  <ADJP>IL-1- <ADJP NULL="QSTN"/></ADJP>
  and
  <ADJP>IL-18-mediated </ADJP></ADJP>
function
</NP>
```




**[図 3.4.4-a] 等位接続 “IL-1 and IL-18-mediated function” の構造化**
**＊印は空要素(“mediated”に対応するものが省略されている)を表す。**

3.4.4 -a は「IL-1 mediated かつ IL-18 mediated であるような function」という解釈であり、図 3.4.4-b は「IL-1 mediated function および IL-18 mediated function」という解釈に相当する。このどちらが正しいかの判断は、言語学を専門とするアノテータにはつけづらい。逆に、生物学の専門家には構文構造に関する知識が不足しているため、構造付け全体の作業が困難なものとなる。今後、このような場合の構造化にいかに専門家の知識を取り込むか、たとえば、意味がわからないため構造があいまいになる部分のみを提示してあいまい性を解消するのに必要な最低限の構造を専門家に付与してもらいその解釈を尊重した全体の構造を言語学者が付与する、などの協力体制を確立することが、生命科学のみならず高度に専門化された内容を表すテキストのタグ付けのために必要となろう。

```
<NP SYN="COOD">
<NP><ADJP>IL-1- <ADJP NULL="QSTN"/></ADJP>
        <NP NULL="RNR" ref="i20"/></NP>
and
<NP>IL-18-mediated <NP NULL="RNR" ref="i20"/></NP>
<NP id="i20">function </NP>
</NP>
```




［図 3.4.4-b］図 3.4.4-a と同じ句に対する別の構造化
＊印は空要素(“mediated”に対応するものが省略されている)を表す。
＊20 印は “function”と対応する空要素を表す。

### ⑤ GENIA 相互作用イベントコーパス

相互作用イベントコーパスは、文中で生体内の相互作用イベントがどのように表現されているのかを明示的に記述することを目的とし、専門用語コーパスの対象アブストラクトのサブセットである 500 アブストラクトに対して、Caderige プロジェクト（フランス）にて定義された DTD を拡張した GENIA-event DTD を用いて XML 形式で付与したものである（図 3.4.5)。

```
<sentence id="U93136418:7" title="no">

    <genic-interaction regulation="inhibit"
type="protein-expression">
    Finally, <exf>using <ex>electrophoretic mobility shift
    assays</ex></exf>, evidence was obtained that <gaf1>
    <ga1 role="modulate" type="protein">IL-4</ga1></gaf1>
    <if><i>inhibits</i></if> <gtf1>LPS-induced expression of
    <gt1 type="protein">AP-1</gt1>
protein</gtf1>.</genic-interaction>

    <non-genic-agent-interaction type="protein-expression">Finally,
    <exf>using <ex>electrophoretic mobility shift assays</ex></exf>,
    evidence was obtained that IL-4 inhibits <ngaf1>
    <nga1
type="exogenous">LPS</nga1></ngaf1>-<if><i>induced</i></if>
    <gtf1>expression of <gt1 type="protein">AP-1</gt1>
    protein</gtf1>.</non-genic-agent-interaction>

  </sentence>
```

［図 3.4.5］ GENIA 相互作用イベントコーパス

相互作用イベントタグは各文章単位で付与されるものとし、文章は相互作用を記述するものとしないものに分類される。相互作用を記述する文章は、“Assertion”、“Regulation”、“Type”、“Uncertainty” 、“Self-contained” 、“Confidence”の６種類の属性を持ち、肯定文か否定文かの極性や相互作用の性質と種類、表現の確実度、情報の完結性、作業者の自信度を表現することができる。それぞれの相互作用は、主に作用者（Agent）、標的（Target）、作用（Interaction）から構成され、他に確実さや時期・時間、場所、実験手法に関する表現などを付随情報として付与することができる。

**［関連発表文献］**

(Kim et al., 2003) Kim, Jin-Dong, Tomoko Ohta, Yuka Teteisi and Jun'ichi Tsujii. (2003). **GENIA corpus - a semantically annotated corpus for bio-textmining.** Bioinformatics. 19(suppl. 1). pp. i180-i182. Oxford University Press.

(Kim et al., 2004) Kim, Jin-Dong, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi and Nigel Collier. (2004). **Introduction to the Bio-Entity Recognition Task at JNLPBA.** In the Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04). pp. 70--75.

(Tateisi and Tsujii, 2004) Tateisi, Yuka and Jun'ichi Tsujii. (2004). **Part-of-Speech Annotation of Biology Research Abstracts.** In the Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004). IV. pp. 1267-1270.

(Tateisi et al., 2005) Tateisi, Yuka, Akane Yakushiji, Tomoko Ohta, and Jun’ichi Tsujii (2005) **Syntax Annotation for the GENIA corpus.** To be presented at IJCNLP 2005, Jeju, Korea, October10-15.

### (2) 研究成果の今後期待される効果

現在、専門用語コーパス・品詞コーパスは 2000 アブストラクト（約５０万語）を公開し、外部の多くの研究者により活用されている。

生命科学分野のテキスト処理システムの開発においては、いまだに多くの場合個々の開発者が、独自の小規模なコーパスを使用してシステムの開発・評価を行っていることから、研究の評価や相互比較が極めて困難になっている。GENIA コーパスは公開以来この分野のコーパスとして最大規模のものとして、新たな手法開発のためのコーパスとしてだけでなく、研究成果の比較を行うための標準的なコーパスの地位を獲得している。

米国・コロラド大学のグループは 2005 年の ISMB 併設 Biolink ワークショップにおける主な生命科学論文コーパスの使用状況を調査し、GENIA コーパスが 21 件と他（10 件未満）に比べて圧倒的に多いことを報告している。現在、米国・フランス・スウェーデンなどの多くのグループが生命科学分野でのコーパス作りを開始しているが、ＧＥＮＩＡコーパスほど網羅的に各レベルの注釈を付加しているものはない。また、ここで報告したもの以外に、シンガポールのグループ（Inforcomm）が共参照関係の付加、ドイツのグループ（EML）がＧＥＮＩＡオントロジーの拡張を行うなど、GENIA コーパスを中核にした国際協力も始まっている。これらの試みと同時に、同種の開発を行っているペンシルベニア大学、ＣＮＲＳとの人材の交流を含めた共同研究も始まっており、ＧＥＮＩＡコーパスとその後継が今後の生命科学分野でのテキスト処理に与える影響は極めて大きい。

## ３．５　情報モビリティ向上のためのソフトウェア基盤の作成
## ［東京大学　ソフトウェアグループ］

### (1) 研究実施内容及び成果

ソフトウェアグループは、大規模なデータに対する、大きな計算量を必要とする言語処理を支援することを任務としたグループである。本グループはそれを実現するために、分散環境における並列処理を簡便かつ効率よく行う方式、ウェブからの大規模テキストデータの高速・柔軟な収集と蓄積、などについて研究を行った。

主な成果として、(a) 分散環境で、日常作業、計算機管理から、並列処理までを効率的に行うための環境・ツールの設計と実装、(b) 言語処理グループとの協調による、大量の MEDLINE アブストラクトの HPSG 構文解析((a)の成果を利用)、(c) ウェブから高速にページをダウンロードするエンジンの設計と実装、それを用いた 2 億ページ/圧縮時 1.3TB 程度のウェブデータの収集、(d) 言語処理グループとの協調による、大量の日本語文の CFG 構文解析((a)(c)の成果を利用)、がある。

これらにより、全体として、大量のウェブコーパスを用い、大規模な計算資源を投入した言語処理を行う基盤ソフトウェアが構築された。(b)・(d)は、統合実験（３－７節）で報告することとし、本節では、(a), (c)について報告する。

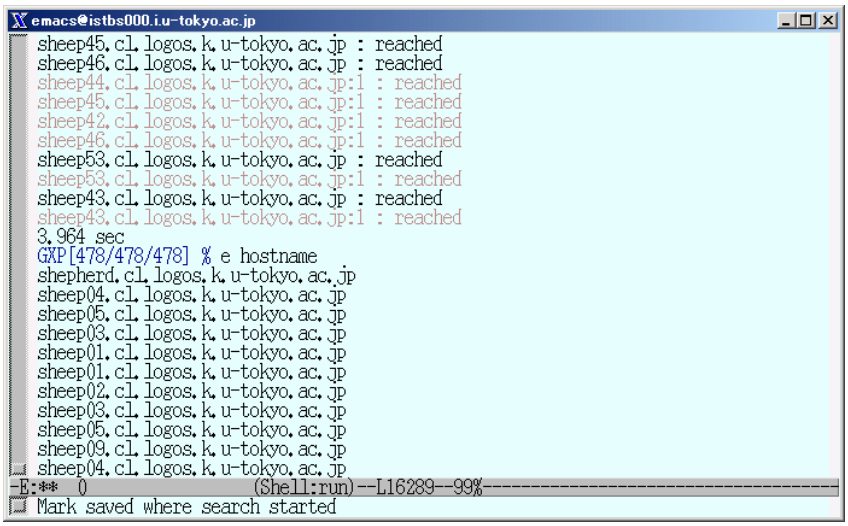
#### ① 分散環境での計算機管理から並列処理までを効率的に行うツールＧＰＸ

現在、高性能計算環境として、数十台から数百台の計算機を 100Mbps～数 Gbps クラスのローカルエリアネットワークで結合したクラスタ計算機が主流となっている。また、そのようなクラスタがキャンパス内あるいは広域に複数存在している分散環境も一般化している。これらは現在、多くの用途に対して、最高の費用対性能比をもたらす計算機環境である。しかしその一方で、これらの環境を並列処理の専門家以外のユーザが簡便に利用するためのソフトウェア環境は整っていない。たとえば分散環境では、ファイルの共有や負荷分散など、単一システム環境やクラスタ環境で、(部分的には)オペレーティングシステムによって自動的に提供されていた機能が欠落している。このことは、特に分散環境において、プログラムの開発、並列化、インストールなどのコストを著しく増大させる。個々の問題点を解決するソフトウェアは多数提案されているが、逆にそれらのソフトウェアを導入、維持するためのコストが多大であったり、管理者権限を必要とするためにすべての計算機に共通する環境を提供できない場合が多い。これらによって多くの潜在的利用者が分散環境から遠ざかっているのが現状である。

我々が設計・実装したツール GXP は、通常のシェルを自然に延長したインタフェースで、しかし多数の計算機を同時に使うことを基本アイデアとしたツールである(図 3.5.1)。図の中段の青字 GXP[478/478/478]がシェルと同様のプロンプトで、ここに続いてコマンドを入力する。数字の 478 は、現在 478 台のプロセッサ(正確には、それぞれ 2 プロセッサを持つ 239 ホスト)を同時に操作していることを示している。

“e hostname”は投入されたコマンドで、これにより全ホストで一斉にコマンド hostname が実行され、引き続く行に、全プロセスからの出力が結合されて表示される。

実行は非常に迅速で、上記の環境で、起動から終了までを 0.5 秒以下で行う(後述)。
ホストは東京大学本郷キャンパスと、柏キャンパスに設置されている(図 3.5.2)。




[図 3.5.1] GXP コマンドラインインタフェース

本郷キャンパスクラスタ

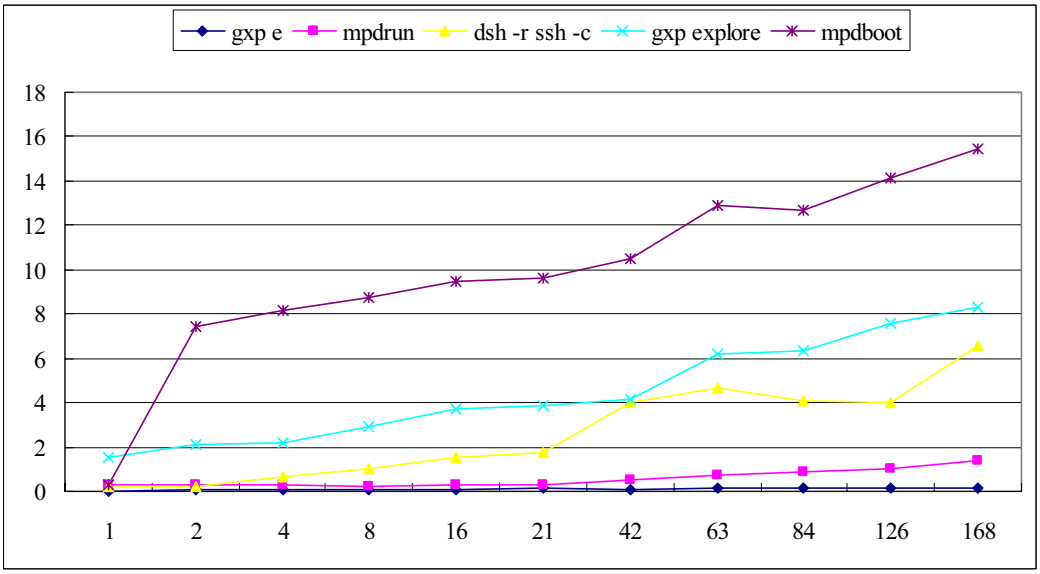
柏キャンパスクラスタ

[図 3.5.2] GXP ホストコンピュータ

GXP と類似のツールとして、dsh や pdsh などの、ssh や rsh を用いた遠隔コマンド起動を並列に行うツール、mpd や pvm などの高速な並列プログラムランチャーがある。それらと GXP は、以下のような数多くの点で異なる。

1. 多数ホストに対する高速なプロセスの起動：
   GXP は既存のツールと比べて、台数が多い(100 台を超える)環境で、一桁から二桁高速である。それには、既存のツールの多く(dsh, pdsh)はコマンド実行のたびに認証を含めたリモートログインを実行しているのに対し、GXP はいったん生成された接続を維持して、以降のコマンド実行を高速化していることが大きく寄与している。高速な並列プログラムランチャーである mpd は GXP と同様の構造をしているため、

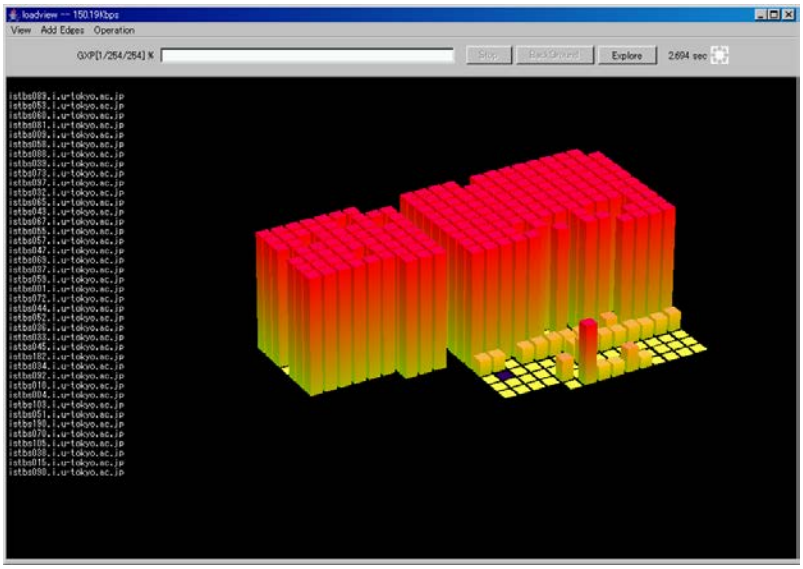
dsh や pdsh よりも高速であるが、プロセスの起動のための通信にリング構造を用いている。一方 GXP は木構造を用いているため、たとえば 150 台を超えるような大きな台数では、GXP は mpd より一桁程度速い。図 3.5.3 はそれらのツールの性能比較を示している。空プロセスを実行するのにかかる時間を比較しており、下へ行くほど高速である。




[図 3.5.3] 関連する研究(類似ツール)との性能比較。
空プロセスを起動するのにかかる時間(縦軸:単位は秒)を台数(横軸)を変えて測定したもの。大きな台数(168 ノード)GXP のプロセス起動時間(gxp e)は次に高速な mpd (mpdrun)と比較しても一桁高速であり、対話的なレスポンスを提供している。

2. 導入を容易にすることを最初から念頭に置いた設計:
   GXP は複数の、ファイルシステムを共有していない環境で、最小限の初期設定で動くことをはじめから設計の目標にしている。最初に GXP を起動する一ホストへインストールすると、そのほかのノードへは自動的に GXP がインストールされる。また、クラスタ内に存在するホスト名の取得なども自動的に行われる。全体として、初めて使うユーザが数分のうちに、リモートログイン可能な数百のノードの同時利用を開始できるように設計されている。既存のほとんどのツールはすべてのノードで、そのコマンドが同一パスにインストールされていることを前提条件にしている。また、コマンドを実行するホストをすべて列挙することもユーザの手間となっている。
3. 分散環境の多様性、複雑なネットワーク設定を考慮した設計:
   GXP は複数の管理ドメイン(LAN)にまたがった、多様、かつところによってネットワーク接続が制限された環境で動作するように、最初から設計されている。既存の多くのツールは、ユーザ名がすべてのノードで同じであること、どのノード対間でも直接ログインが可能であることなど、実質的に、すべての計算機が一つの管理ドメイン(LAN)内にあることを仮定している。また、GXP はインストールに管理者権限も必要なく、利用者が、自分がアカウントを持つ計算機すべてを利用するための障害を極力少なくするように設計されている。

GXP は http://www.logos.ic.i.u-tokyo.ac.jp/phoenix/gxp_quick_man_ja.shtml にて公開している。図 3.5.4 は、分散環境へのアクセスをより容易にする GUI である。




[図 3.5.4] GXP GUI。各ホストの負荷(CPU 使用率が、バーと色で示されている)

### ② 柔軟な WWW データ収集のための高速ダウンロードエンジン

WWW のデータを利用する研究は数多く行われているが、その多くは WWW 上で稼動している検索エンジンへ検索要求を行ってデータを入手している。この方法で、現実的な時間内で入手できるデータには限界があり、また情報検索や抽出の目的、用いるアルゴリズムに関係なく、常にキーワードによって取得したいデータの絞込みを行わなくてはならないなど、制限も多い。そのため、WWW からの大規模な情報抽出やコーパスの作成などのためには、自力でページのダウンロードや索引付けを行う必要がある。もちろん既存の検索エンジンなどでも同様の機構を実装しているが、全データを直接アクセスできるように公開されているわけではない。また、WWW のデータ量は膨大であるため、そもそもどのようなデータを中心に収集すべきかも、目的とする研究によって異なり、柔軟で高速な収集エンジンを持つことの意義は大きい。

本研究では、ウェブからの情報抽出を支援するためのツールとして、高速で安全な収集エンジンを設計・実装した。研究期間前半に第一版の実装を行い、後半に再設計と再実装を行った。我々の設計したエンジンは単純なコマンドラインインタフェースを持ち、標準入力に与えられた URL を持つページを順不同で収集し、結果を、ページごとに区切りを入れて標準出力に出すというものである。これを、得たページの解析を行うプログラムや、online に入力ページを計算するプログラムとパイプで結合したり、GXP を用いて並列に起動するなどして、柔軟な収集器が容易に構築可能である。収集エンジンはいわゆる WWW 上のエチケット(特定のサーバに集中して負荷をかけない、など)を守る必要があり、高速な収集のためには多数(1,000 近く)のコネクションを維持して多数のホストから少しずつ並行してダウンロードを行う必要があり、我々のプログラムはこれ

を行って、1 台のホストで 500 ページ/秒、20 台で 10, 000 ページ/秒の速度を達成している。単純計算で外挿すると、1 日に 8 千万ページほどの収集能力がある。ただしこれは短時間の記録であり、今後長時間の運用を行うとともに、収集したページからのリンク抽出や索引付けなどを含めた総合的な評価と有用性の検証を行っていく必要がある。

研究期間の前半に、1 ヶ月程度で約 2.5 億ページ程度を収集した(圧縮時 1.3TB)。このデータは、言語グループ(西田グループ)の黒橋、川原らによって、ウェブコーパスからの格フレームの学習のためのデータとして用いられた。2.5 億ページはデータを収集したクラスタ内の 4 ノードからなるファイルサーバにまたがり、それを上述した本郷・柏に設置されたクラスタの約 350CPU で処理している。各ページに対して、ファイルサーバからクラスタへのデータ転送、言語認識(日本語ページの抽出), JUMAN による形態素解析、KNP による構文解析を実行している。並列処理にはやはり GXP を用いており、概算で、1,000CPU 日程度の処理を、3.5 日程度で処理している。

### (2) 研究成果の今後期待される効果

日々生産、蓄積されていく WWW 上の膨大なテキストに対して、ルーズに統合された大量の計算資源を用いて、実際に深い自然言語処理を行うことができることを実証したことに意義がある。我々の経験によれば、これらの処理を実行するために、人的コストの高いミドルウェアやソフトウェアインフラの整備、管理者の協力はほとんど必要がなく、いわゆる LAN につながった通常の設定の Unix 計算機のネットワークで事足りる。

本研究項目で作成されたＧＸＰは、３－７節の統合実験に使われ、処理できるテキスト量を２桁以上向上させることに成功し、今後の大規模テキスト処理への基盤技術を確立した。

言語処理今回の実験で我々は、負荷分散のためのシステムや、分散環境のためのファイルシステムなどの特別なミドルウェアは用いていないが、それらの機能は GXP が代わりを提供しているか、GXP があれば(今回の実験の範囲内では)不要といえるものであった。長時間にわたるジョブの実行は、我々のジョブの優先度を最低にした状態で、ほかのユーザと通常通り資源を共有した状態で行われ、我々が計算機から他のユーザを排除したわけではないことも重要である。これは、並列計算が研究の道具であって、目的ではない研究者には朗報であり、廉価なハードウェアに加えて、廉価な人的コストで、本題の実験に時間を割くことを可能にし、巨大な計算能力を日常的に使用できる環境を構築できることを示した。この成果は、テキスト処理・言語処理に限らず、広い科学技術の研究での計算機環境を構築するのに有効な手段となる。

言語処理は、今後、大きな辞書・オントロジーに基づく知識処理と言語処理とを統合する、より大きな計算能力を必要とする研究に進むことは必至である。今回の研究で構築した基盤の上に、分散環境下でのファイルシステムを構築する技術を開発することで、次世代のテキスト・言語処理のためのソフトウェア基盤が構築できると考えている。

## ３．６　身体性を考慮したエージェントとの実世界インタラクション方式
## ［京都大学　エージェントグループ］

### (1) 研究実施内容及び成果

本グループでは、実世界環境において身体を用いた非言語表現によって感情・社会的なインタラクションを行う実世界会話エージェントの開発を行い、実世界における情報のモビリティ促進に寄与する技術を研究開発する。

とくに、ロボットをコミュニケーションのためのメディアとして位置づけ、人間とロボットの身体的なインタラクションによって、感情的なコミュニケーションと意思疎通を行う技術の開発を目指した。これは、テキストという言語媒体による情報の流通を円滑にする言語処理グループと研究と相補的な視点を持つもので、将来、この２つの情報伝達モードを統合するための枠組みの基礎となる。研究成果として、引き込み原理と確率推論を使った人間・人工物インタラクション方式という基本的な枠組みを提案し、その有効性を実装システムで実証した。引き込み原理を使ったインタラクション方式は、個々のジェスチャの意味づけを予め行わないノンシンボリックジェスチャによるインタラクションを世界に先駆けて実現したものである。

#### ① 引き込み原理によるインタラクションとその応用

引き込みとは、近いリズムで振動しているシステムの挙動が同期する現象である。引き込みは、人のコミュニケーションでも重要な役割を果たしていると考えられる。話がはずんでくると知らず知らず同じリズムで頷きあっているというのはその一例であり、シンクロニー（「同調」）として知られている。

引き込み原理を用いることによって、予めシンボリックジェスチャによるインタラクションのプロトコルを定めておかなくてすむノンシンボリックジェスチャによるインタラクション方式の実現が期待できる。

本研究では、引き込み原理を使ってインタラクションに参加している行為者間の行為の相互調整を行うシステムを開発した。このシステムは、同調と変調というフェーズを経て行為の相互調整を行う。同調のフェーズでは、ユーザが繰り返し動作により指示を行うと、ロボットがそれに対応したアトラクタを持つ力学系を内部に構成する。次の変調のフェーズでは、ユーザが自分の繰り返し動作のリズムを微調整すると、ロボットもそれに応じて力学系を修正し、自らの動作を変える。同調と変調を繰り返すことによって、ロボットの持つ様々な機能を引き出し、ユーザの目的を達成できる。

引き込み原理を使ったインタラクション方式を検証するために床掃除ロボットシミュレータを研究開発した。大型ディスプレイ上には、シミュレートされた部屋が表示される。ユーザは、モーションキャプチャを取り付けた右手や左手の繰り返し動作によって、大型ディスプレイ上に表示された複数台の床掃除ロボットと家具運搬エージェントに対して、シミュレートされた部屋をどのように「掃除」するかを指示する。床掃除ロボットへの指示はそれぞれのロボットに個別に出す場合と全体に指示を出す場合の両方がある。片手で指示を行うときはその手に対応する指示ポインタの位置から一番近く

にいる床掃除ロボットがユーザの指示に反応する。右手と左手の指示動作がほぼ同一であるときは、指示は全体命令と解釈され、すべてのロボットがユーザの指示に反応する。

床掃除ロボットの基本的な動作のモードは、自律動作モジュールによって自律的に動作する場合と、自律動作モジュールによる制御を受けず、相互調整モードで動作する場合に分かれる。自律動作モードでは、各床掃除ロボットは、限定された知覚能力と知能により、ユーザからの指示なしで障害物を回避したり、近くにある汚れた場所を感知して掃除をしたりすることができる。床掃除ロボットは、通常は自律動作モジュールに従って動作するが、繰り返し動作といったユーザの意図的な検出すると相互調整モードに移行する。

床掃除ロボットは、ユーザの右手と左手の位置データ $(X,Y,Z)$ を常に取り込み、それを２次元データ $(X,Y)$ に投影して、指示ポインタの位置データ系列として一定期間記憶する。記憶した位置データ系列長が、予めパラメータとして設定した同調ウインドウサイズ $W$ 時間分に達したところで、同調が起きているかどうかの判定を行う。まず、得られたデータ系列から直流成分とノイズを除去し、データ系列が原点を中心とする半径１の円内に収まるように正規化を行う。このとき、当該の床掃除ロボットの内部状態と位置の間の対応関係も決定される。次に、データ系列の自己相関関数を計算し、ピーク値の時間遅れを周期 $T$ とする。ピーク値が閾値を越えていたら同調のための繰り返し動作が行われたものと判定し、繰り返し動作のパターンを計算する。データ系列を周期 $T$ ごとに分割し、平均化して１周期分（$T$ 時間）のデータ $(x,y)$ の系列を抽出する。離散フーリエ変換と離散逆フーリエ変換を行い、$l$ 個の低周波成分を抽出する。これにより、繰り返し動作のパターンとして、ノイズが除去された周期 $T$ の閉軌道を表す式を得る。次に、アトラクタ周辺の点について、その点を通る軌道がその点に近いアトラクタ上の点 $(x,y)$ に近づくように力学系を構成する。

変調は、同調によって検出された繰り返し動作のパターンのリズムを基調とし、それに一定の差分（位相のずれと周期比）を定常的に加えつづけることによって生じる動作である。

まず、同調によって得られた当該ロボットの内部状態の遷移を支配する力学系を固定して、当該ロボットの内部状態と位置との間の対応を求める。

次に、ユーザの繰り返し動作から変調ウインドウサイズW時間の画面上の指示ポインタ位置のデータ $(X,Y)$ を抽出する。このデータとアトラクタデータを比較し、写像関数を求める。データ群から最小二乗法により、線形トレンド $ct+d$ を除去する。次に、自己相関関数により周期 $T'$ を抽出し、アトラクタ周期 $T$ との比 $m=\frac{T}{T'}$ を算出する。そして、周期 $T'$ に変換したアトラクタの式

$$x=A(mt),y=B(mt)$$

と相互相関関数を計算し、ピーク値から時間遅れ n を検出する。こうして得られたアトラクタデータ

$x = A(mt+n), y = B(mt+n)$

を変調後の動作の基本リズムとする。

最後に、$(x,y)$を座標$(X,Y)$に変換するための写像

$X = P(x,t), Y = Q(y,t)$

を求める。

上に述べた引き込み原理に基づく相互調整をもちいた床掃除ロボットの性能を、リモコンによって完全に手動で掃除機を動かす場合や、完全自動掃除アルゴリズムによって掃除機を駆動する場合と比較した結果、相互調整をもちいて掃除ロボットに指示を与える場合は、手動のようにつききりで床掃除ロボットを制御しなくても、完全自動掃除アルゴリズムより高い性能が得られることがわかった。

**② 確率推論を用いた人間とロボットのインタラクション**

ロボットが人間とより複雑なインタラクションができるようにするためには、人同士のコミュニケーションの作法に近いモデルに基づいて人とコミュニケーションできるようにする必要がある。本研究では、コミュニケーションの開始において場を確立するプロセスに焦点をあて、ロボットが客に近づくにつれて、「客の注意を自分に向ける」、「客にコミュニケーションしたいという意図を伝える」、「向き合う」などの行動をとれるようにすることをめざした。

ロボットのコミュニケーション行動を生成するための手がかりとなる非言語情報としては、対人距離、視線情報、ACK(手を挙げる、頷きなどの確認動作)の３つを用いることとした。ロボットが人とインタラクションするときのパターンを、腕の動作、視線動作、接近動作などの系列とその動作を実行するための条件(視線、対人距離、ACK) を合わせてパッケージ化したインタラクションシーンとして、複数個のインタラクションシーンをインタラクションの流れに従って並べたインタラクションスキーマとして、それぞれ表現する。

インタラクションシーンの集まりとして規定された行動知識に基づいてウェイターエージェントの行動を決定するために、インタラクションシーンの各要素の値をセンサー情報・行動履歴と関連づける。センサーから得られた情報と行動履歴の情報がインタラクションシーンの各ステップに供給され、このインタラクションシーンの妥当性が恒常的に点検される。現在のインタラクションシーンの適用が合理的でないと判断されたときは、別のスキーマへの遷移が行われる。適用可能と判断された場合は、「視線が一致」、「ACK の交換」など重要な非言語コミュニケーション記録が行動履歴管理のためのデータベースに記録される。こうした総合的な判断によって、センサ情報が不完全で限定的なものであっても、ロボットは行動プランをたてることができる。

ロボットは入力されてくる観測データに基づいてインタラクションシーンと現在の状況を継続して照合することによって、客とのインタラクションが典型的なパターン通りでない場合でも、インタラクションシーンを切り替えて柔軟に対応できる。そればかりか、インタラクションシーンに基づいて次のユーザの行動を予測していることにもな

るので、ユーザの行動を予測して先んじて行動することもできる。

実際の物理環境では、ロボットに備わっているセンサでは客の全ての動作を認識することができない。画像認識やモーションキャプチャなどによって人の動作を認識しようとしても、情報不足やノイズのために人の動作のほんの一部分しか認識しかできない。さらに、人が気まぐれな行動をすることによる不確実性が生じる。そのような場合でも、観測された不足情報の一部分からロボットの挙動を迅速に決定する必要がある。こうした状況に対応するために、動作を確率的に表現することとした。確率推論にはベイジアンネットワークを使った。

上に述べた方式を、Robovie を用いて実装し、有効性を確認した。

**[関連発表文献]**

安西祐一郎、山崎信行、徳田英幸、西田豊明、萩田紀博、廣瀬通孝. ロボットインフォマティクス. 岩波講座ロボット学５. 岩波書店. 刊行予定.

Takashi TAJIMA, Yong XU, Toyoaki NISHIDA. **Entrainment Based Human-Agent Interaction**. To be presented at 2004 IEEE Conference on Robotics, Automation and Mechatronics, December 1-3, 2004. Traders Hotel, Singapore.

畠山誠、西田豊明：同調動作に基づくロボットと人間のコミュニケーション、 第 17 回人工知能学会全国大会、 1D1-05、 2003.

畑田寛久、畠山誠、西田豊明：ロボットと人とのコミュニケーションの誘進的確立、第 17 回人工知能学会全国大会、 1D1-06、 2003.

西田豊明：人とロボットの意思疎通、人間とロボットの意思疎通、特集「情報科学の総力をあげたロボット技術」情報処理、 44 巻、12 号、pp. 1214-1220. 2003.

### (2) 研究成果の今後期待される効果

知的な情報検索システムが有効に機能するためには、検索結果の視覚的な提示などを、従来のテキストによる提示（たとえば、抄録）と結合することの必要性が認識され始めている。しかし、システム側からの情報提示の方式の研究に対して、ユーザが非言語的な手段を用いて意図を表現するための系統的な研究は、既存のＧＵＩ研究の延長で考えられてきた。今回の研究では、既存のＧＵＩに比べてはるかに動的な情報伝達ができる引込みにより情報伝達という視点を提示したことで、言語によるコミュニケーションを補完する非言語的な情報伝達の可能性を示した。

今後、この方向での研究が、言語的なものによるコミュニケーションと融合されることで、ユーザ指向型の情報システムの研究がより豊かなものになることが期待できる。

## ３．７　統合実験

［東京大学　ソフトウェアグループ，東京大学　言語処理グループ］

### (1) 研究実施内容、および、成果

プロジェクト後半は前半で研究された理論と要素技術を統合することで、実世界での応用に耐える高性能で高精度なシステムを開発することを目指した。３－１節の英文解析器 Enju は HPSG の高速化、コーパス指向の文法開発という技術だけでなく、３－３節、３－４節の研究成果を統合することで実現したものである。さらに、本節の統合実験では、(a)言語処理グループが開発した Enju を、(b)ソフトウェアグループが開発したＧＰＸ（３－５節）を使うことで、生命科学分野の文献の巨大な集合に適用することで、言語処理グループ、ソフトウェアグループの研究を統合することで、我々の開発してきた要素技術の有効性を示すことを目的とした。

文解析は、３－８節の総合システムを構築するための不可欠な技術であり、それが実用上問題のない時間内で実行できることを示すことは、本研究プロジェクトの最重要課題である。

文解析の結果が情報抽出の精度を著しく向上させること（Yakushiji et al., 2005; Chun et al., 2006)、Enju による解析精度が世界最高水準のものであることはすでにわかっているので、統合実験実験では処理時間に焦点を当てた。

以下に実験の概要を示す。

#### ① MEDLINE

MEDLINE とは、NCBI (National Center for Biotechnology Information) が提供する文献データベースである。医学･生物学に関する論文のほとんどをカバーしており、2005 年時点で、収録論文数は約 1500 万件に達する。パーズの対象は、タイトルおよびアブストラクトの中身とした。表 3.7.1 にパーズ対象とした MEDLINE コーパスの統計情報を示す。収録論文全体のうち、アブストラクトが利用可能であるものは約半数であることがわかる。また、１つのアブストラクトあたりの平均文数は約 10 文、１文あたりの平均単語数は約 20 単語である。総単語数は約 14 億単語に達し、自然言語処理でよく利用される Penn Treebank の単語数が 100 万語であることを考えると、きわめて大規模なコーパスであるといえる。

| | |
|---|---|
| 論文数 | 14,792,890 |
| アブストラクトが存在する論文数 | 7,434,879 |
| 総文数 | 70,815,480 |
| 総単語数 | 1,418,949,650 |

［表 3.7.1］ MEDLINE コーパス

#### ② HPSG パーザ（構文解析プログラム）

MEDLINE の構文解析には、Enju を利用する。Enju は Penn Treebank をもとにして構築された高被覆な HPSG 文法を利用しており、さらに②で解説されているさまざまな高速化技術を利用することで、高精度かつ高効率なパージングを実現している。

オリジナルの Enju は Penn Treebank という新聞記事をメインにしたコーパスを利用して文法学習がなされており、それをそのまま MEDLINE という生物医学テキストに適用すると、語彙の違いや構文的な preference の違いにより精度が大きく低下する。そこで、本グループでは、GENIA コーパスを利用して、学習モデルのパラメータを生物医学テキスト用に最適化する（Hara et al., 2005）。表 3.7.2 に最適化前後の GENIA コーパス上での構文解析の精度を示す。パラメータの最適化により精度が大きく向上していることがわかる。また、この最適化においては、Penn Treebank で学習したモデルを、最大エントロピー法における事前分布として利用するという手法を用いており、非常に小さな計算コストで新たなドメインのテキストに最適化できるという特長がある。

| | Penn Treebank | GENIA Treebank |
|---|---|---|
| オリジナル Enju | 87.57 | 84.57 |
| MEDLINIE パーズ用 Enju | 87.31 | 86.02 |

［表 3.7.2］ HPSG パーザの精度（F-score）

同様に、品詞タガーに関しても、MEDLINE 用にチューニングしたものを利用する (Tsuruoka et al., 2005)。表 3.7.3 に学習用データとして複数のコーパスを利用した場合の結果を示す。GENIA コーパスと PennBioIE コーパスはどちらもバイオ分野のコーパスである。今回の MEDLINE のパーズには表最下段の３つの学習コーパスすべてを利用して学習を行ったタガーを使用している。

| | WSJ | GENIA | PennBioIE |
|---|---|---|---|
| WSJ | 97.05 | 85.19 | 86.14 |
| GENIA | 78.57 | 98.49 | 86.59 |
| PennBioIE | 85.45 | 93.20 | 97.74 |
| WSJ + GENIA | 96.96 | 98.32 | 91.98 |
| WSJ + PennBioIE | 96.94 | 93.34 | 97.75 |
| GENIA + PennBioIE | 85.60 | 98.35 | 97.63 |
| WSJ + GENIA + PennBioIE | 96.89 | 98.20 | 97.68 |

［表 3.7.3］ 学習コーパスと品詞タガーの精度（%）

HPSG パーザはスタンドオフ形式でパーズ結果を出力する。スタンドオフ形式とは、アノテーション情報をテキスト本体とは別に保持する形式であるが、表現する内容自体は xml と等価である。パーザから出力されるアノテーション情報は、句構造と述語・項構造である。句構造に関しては、xml でテキスト上にそのまま表現できるために問題はないが、述語・項構造については、句をまたがる遠距離の依存情報を表現する必要がある。そこで、述語・項構造に関しては、それぞれの句に対して一意の id を付与し、係り先を id で示すことで、句の境界を越える依存情報を表現している。

### ③ 並列処理

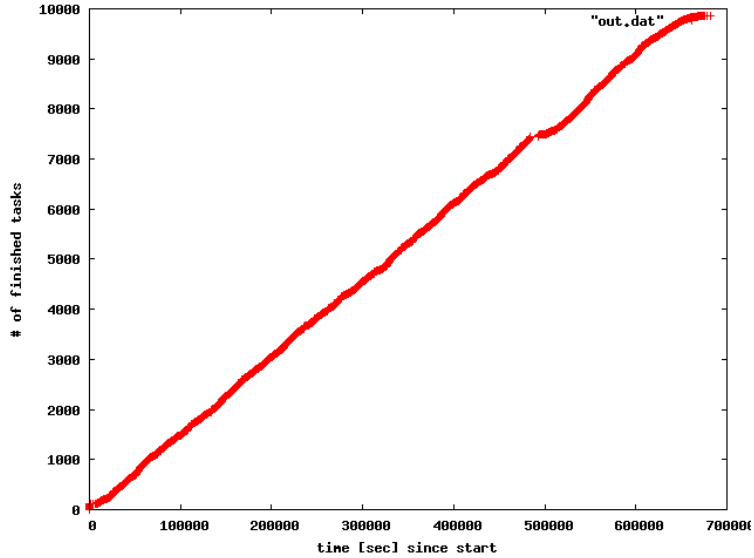
Enju が 1 文の構文解析に要する時間は、AMD Opteron プロセッサ（2.4GHz）上で約１秒である（Ninomiya et al., 2005）。曖昧性解消を高精度で行う HPSG パーザとしては非常に高速であるが、それでも MEDLINE 全体を 1 台のプロセッサで解析しようとする

と、約 7000 万秒、すなわち２年近くの時間がかかることになる。したがって、このような大規模コーパスを構文解析する場合、大量のプロセッサを利用して並列にパーズを行う必要がある。

このために用いた計算資源は前述した本郷キャンパスと柏キャンパスに配置されたクラスタ(合計約 250 ノード中、約 170 ノード/340 CPU を利用)である。並列処理は GXP を用いて行った。アブストラクトは 9,870 個の束に分割され、データサイズは合計で圧縮時 3.2GB 、展開時約 100GB である。二つのクラスタ間に共有ファイルシステムはないため、ファイルは明示的にコピーする必要がある。両クラスタ間は少なくとも 100Mbps 程度でのデータ転送が可能であるため、すべてのファイルをコピーしても数分であり、予測計算時間と比べて非常に短い時間しかかからない。そこで単純のためにすべてのデータを両クラスタのファイルサーバノードに、事前にコピー、配置した。これにより、両クラスタのどのノードでも、すべてのアブストラクトにアクセスでき、負荷分散が単純になる。

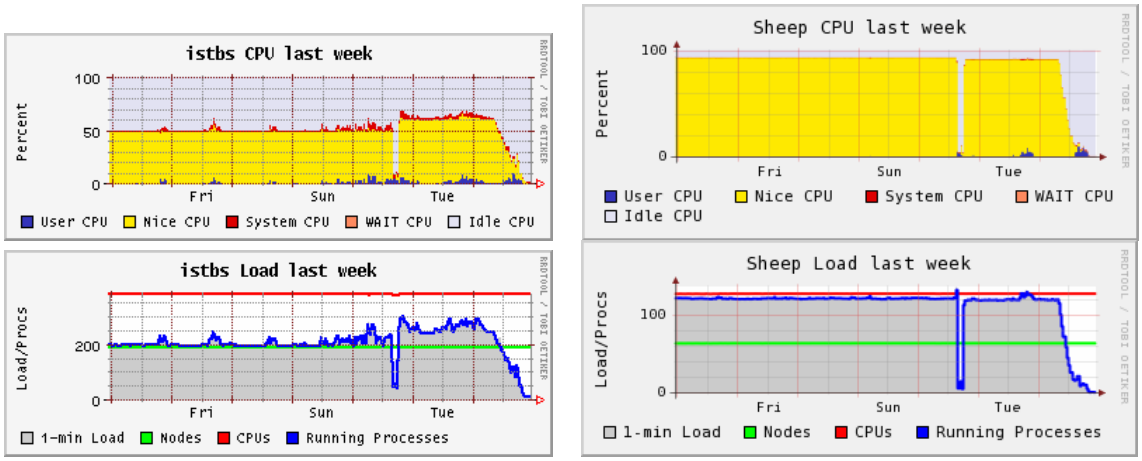
実際の並列処理および負荷分散は、GXP に付随している単純なタスクディスパッチャコマンドを用いて行われた。これは、実行したいコマンドをすべて、ひとつのファイル内に記述しておけば、それらをあいているホストに割り振って実行するというもので、いわゆるバッチスケジューリングシステムと類似の機能を提供する。そして、やはり GXP 以外のソフトウェアのインストールや設定を一切必要とせず、LAN をまたがった資源を容易に統合可能である。終了したジョブの記録はファイルに書き込まれるため、途中で計算機の故障やネットワークの分断などが起こっても、単純に同じコマンドのリストを GXP に渡して実行させることで、計算の続き(終了していないコマンド)の実行が可能である。

実行は 2005 年 9 月 21 日 1:21AM に開始され、同 9 月 29 日中にすべての処理が完了した。この間すべてのプロセスは最低の優先度(nice 値 19)で実行されており、そのほかのユーザもシステムを共有利用していた。




［図 3.7.1］実験開始から終了間際までの経過時間と、終了した束の数

図 3.7.1 は実験開始からの時間経過とともに終了した束の数を示している。図 3.7.2 は本郷キャンパスのクラスタ(左)、柏キャンパスのクラスタ(右)の、それぞれのクラスタ全体の CPU 使用率(上)、ロードアベレージ(下、青線)を示している。横軸の範囲は本実験が行われていた期間約一週間である。途中(右から約 1/3)一度、すべてのプロセスを強制終了しており、その間(数時間)のロードが下がっている。また、本郷キャンパスでところどころロードアベレージや CPU 使用率に盛り上がりが観察されるが、これは他のユーザのジョブによるものである。




［図 3.7.2］CPU 使用率とロードアベレージ

**［関連発表文献］**

(Yakushiji et al., 2005) Yakushiji, Akane, Yusuke Miyao, Yuka Tateisi and Jun'ichi Tsujii. (2005). **Biomedical Information Extraction with Predicate-Argument Structure Patterns**. In the Proceedings of the First International Symposium on Semantic Mining in Biomedicine. pp. 60—69

(Hara et al., 2005) Hara, Tadayoshi, Yusuke Miyao, and Jun'ichi Tsujii. (2005). **Adapting a Probabilistic Disambiguation Model of an HPSG Parser to a New Domain**. In the Proceedings of IJCNLP 2005. (to appear)

(Tsuruoka et al., 2005) Tsuruoka, Yoshimasa, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou and Jun'ichi Tsujii. (2005). **Developing a Robust Part-of-Speech Tagger for Biomedical Text**. In the Proceedings of the 10th Panhellenic Conference on Informatics. (to appear)

### (2) 研究成果の今後期待される効果

ＨＰＳＧという、言語学理論の立場からの文法を現実のテキストに適用し、深い意味構造を計算する処理が１４億語という、大きなテキスト集合に対して適用可能であることを示した、世界で最初のものである。これまでは、大規模といわれる研究でも、ShallowParser と呼ばれる浅い処理を数千万語のテキストに適用する程度であった。今回の統合実験は、従来の技術を質・量の両面で画期的なものである。

また、この実験は、このような大規模実験が、計算機環境としては特別な準備をすることなく、日常的な業務として実現できることを示した点も重要である。実際の応用場面では、MedLine 中のすべてのテキストを処理することは、一度実行すればよく、あとは、新たな増分（たとえば、１ヶ月分）の処理を行うだけで十分であり、現在の環境を

使えば、数時間の日常的な業務となる。この点では、我々の研究成果は、すでに現実の応用場面で使えるものとなっている。

我々の主張、すなわち、「テキスト集合に対してあらかじめ一般的な言語処理を施し、それを蓄積することで、より高度な知識処理と結合するべきである」という主張は、この文解析結果を前提に、３－２節の索引構造を使った知的な検索、あるいは、タスクごとの情報抽出技術を開発することで達成される。このような総合システムの構築ついては、次節で報告するが、この統合実験の成功は、我々の野心的な当初の研究構想が現実的に実現可能であることを示したことになる。

現在、MedLine 全体の解析結果をＸＭＬデータとして公開することを予定しており、３－４節の GENIA コーパスとともに、生命科学分野のテキストマイニングの研究を世界的に推進する基礎データとして活用されることを期待している。

## ３．８　統合サービスシステムの構築
## ［研究グループ全体］

### (1) 研究実施内容及び成果

本プロジェクトの方法論的な特徴は、言語処理・オントロジー技術・エージェント技術・ソフトウェア技術という、分野の異なる技術を統合することで、テキスト中の情報をユーザにとって有効な情報をユーザにとって使いやすい形で提示するサービスシステムを構築することである。

このような統合サービスシステムの構築は、要素技術をユーザシステムにくみ上げる統合技術が必要となるだけでなく、ユーザとの緊密な協力関係を結び、かれらの要求を十分に分析する必要がある。

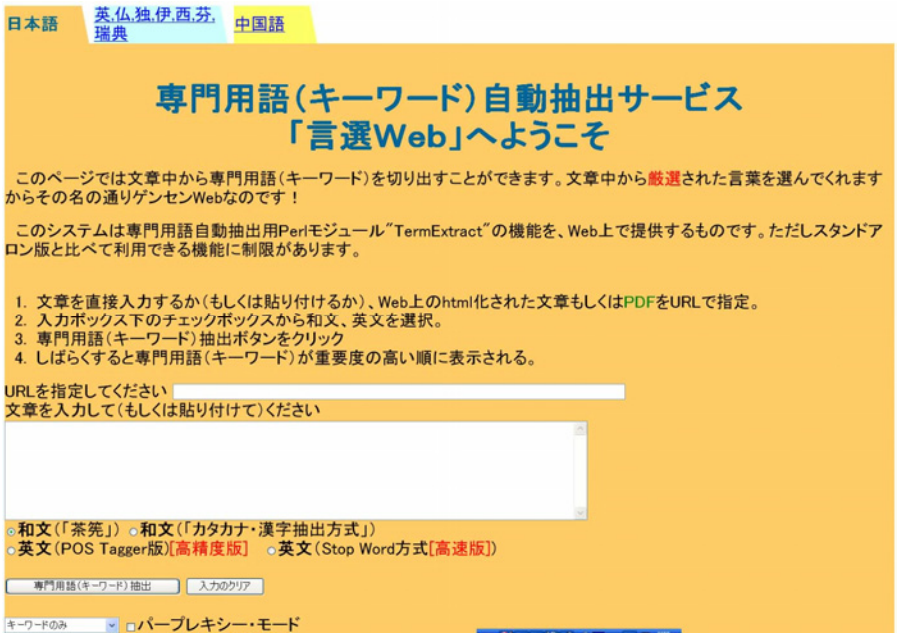
本プロジェクトでは、プロジェクト進行の各時点で可能なユーザシステムを構築し、適宜公開すると同時に、実ユーザとして生命科学分野の研究者集団（特定研究「ゲノムサイエンス」のグループ、産業技術研究所・ＪＢＩＲＣ、など）との共同研究を進めてきた。このような研究の成果として、(a)専門用語抽出サービスシステム、(b)多言語用例検索システム、(c)病疾患・遺伝子関係の自動抽出システムを開発し、一般に公開している。以下では、これらの公開システムについて述べる。

#### ① 専門用語の自動認識システム：言選

オントロジーグループが発展させてきた専門用語の自動認識アルゴリズムを、Web から利用可能なシステムおよびダウンロード可能なモジュールとして整備した。また、同じ手法を日本語以外に英語、その他のヨーロッパの諸言語、中国語に拡大した。このシステムを「言選 Web」という名称で公開した。現在、用語抽出を 3000 アクセス/月、モジュールダインロード 200 件/月である。図 3.8.1 に言選 Web のインタフェースの画面を示す。

分野の専門用語抽出の研究の多くは、分野コーパスを機械学習する方法が採用されることが多いが、言選 Web のアルゴリズムはコーパス依存性、言語依存性を可能な限り排

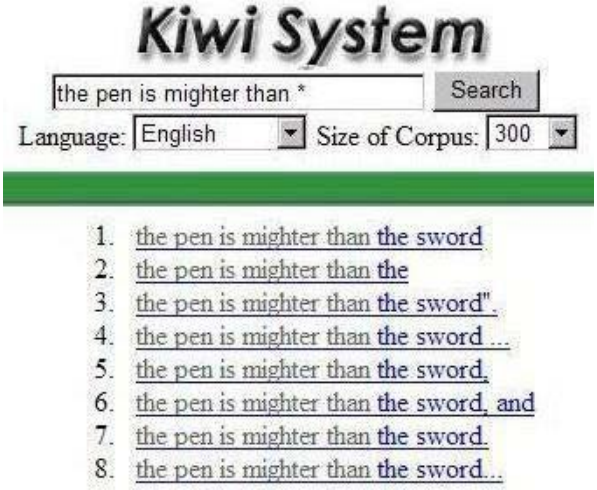
除し、辞書リソースも使わず、場合によっては形態素解析システムさえ省略することが可能であるという簡潔さを身上とする。よって、上記のように多言語への適用は短時間で可能になった。また、この言選システムは、オントロジーグループの研究成果を逐次統合していくプラットフォームとして使われており、現在、単純な手法の専門用語抽出から日本語固有のカタカナ語の取り扱いを行う研究成果などが取り込まれれている。




[図 3.8.1] 言選 Web の日本語対応画面

### ② 多言語用例検索システム

WWW上のテキストを対象にした言語依存性のない文字列ベースの用例検索システムKiwiを構築し、使い勝手、質問応答システムとしとの可能性を評価した。Kiwiの検索画面を図3.8.2に示す。

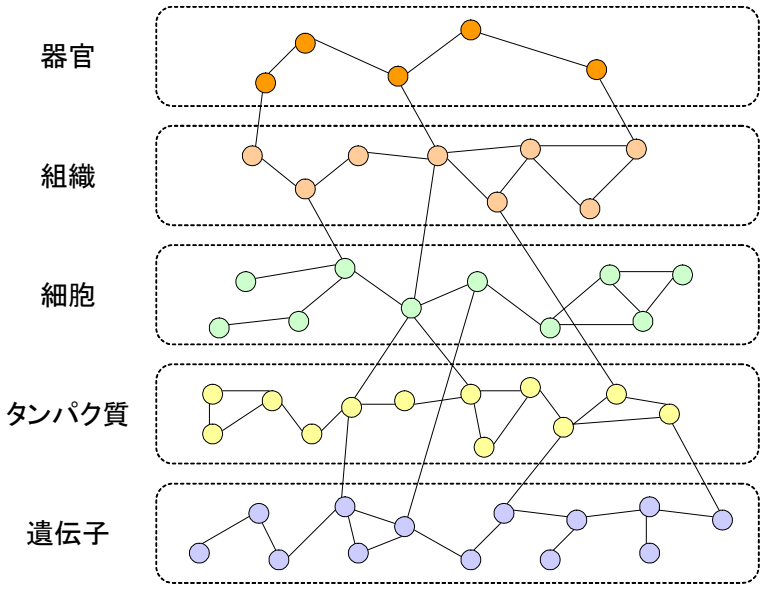



[図3.8.2] Kiwiの検索画面および実行例
（質問は”the pen is mighter than” の後ろにくる言語表現の検索）

同じ仕掛けをローカルなテキストファイルに適用し、知識マイニングのための候補文の絞込み表示を行うPortable Kiwiを試作した。例えば、「による」という表現の前後に現れる文字列をKiwiの手法で集約列挙することによって事象と結果の対からなる知識の候補をマイニングできるようになった。

### ③ 病疾患・遺伝子の関係解析システム

テキストマイニングの役割の一つは、テキスト中の情報を整理・要約された形でユーザに提示することである。医学生物学分野においては、例えば、遺伝子、タンパク質、細胞組織などの相互関係が抽出の対象となりうる重要な知識であり、多くはネットワークの形式で表現することができる（図 3.8.3)。ノードは、遺伝子やたんぱく質といったエンティティに対応し、エッジがそれらの間の関係に対応する。エッジに付与される情報は、それを構成するエンティティのクラスによって異なり、場合によっては、関係の有無だけではなく、関係の種類、positive/negative、confidence などに関する情報が含まれる。




［図 3.8.3］ バイオ知識のネットワーク

我々はすでに（Yakushiji et al., 2005）において、述語・項構造を利用することで、高い精度でタンパク質間相互作用を抽出できることを確認している。また、疾患と遺伝子の間の関係についても、言語処理技術と機械学習を組み合わせることで、再現率をそれほど落とすことなく、高い精度の関係抽出ができることを確認している（Chun et al., 2006)。

このような研究成果を、実際のバイオ研究者が「使える」ようにするためには、抽出された情報を、整理・統合してユーザに提示し、さらに、ユーザの視点や知識を反映した形で得られた知識を整理しなおすことができるシステムを構築する必要がある。

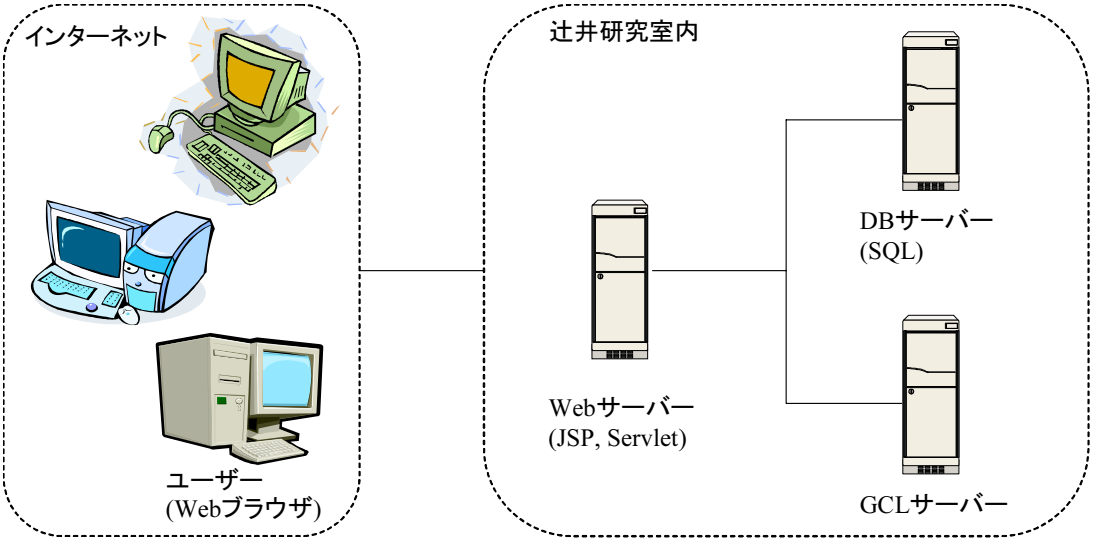
バイオ分野のための既存のテキストマイニングシステムと我々のシステムとの違いは、

深い構文解析の結果を利用することで、従来手法よりも精度の高い関係抽出が可能になっていることにもあるが、GCL および素性構造データベースを利用することで、直接に文の意味構造を指定した検索の機能をユーザに提供できる点にもある。

**システム構成**

統合テキストマイニングシステムの実現方法としては、生物学データベースのインターフェースなどに広く利用されている、Web アプリケーションという形態を利用する。Web アプリケーションは、ブラウザさえ用意できればどこでも使うことができ、すべての機能がサーバ側で提供されているため、機能の修正や変更も容易である。図 3.8.4 に本統合システムの構成を示す。

Web サーバでは、JSP と Servlet によってユーザとのインターフェースを実現している。ユーザは Web ブラウザを通して、情報の検索・保存など、すべての操作を行うことができる。

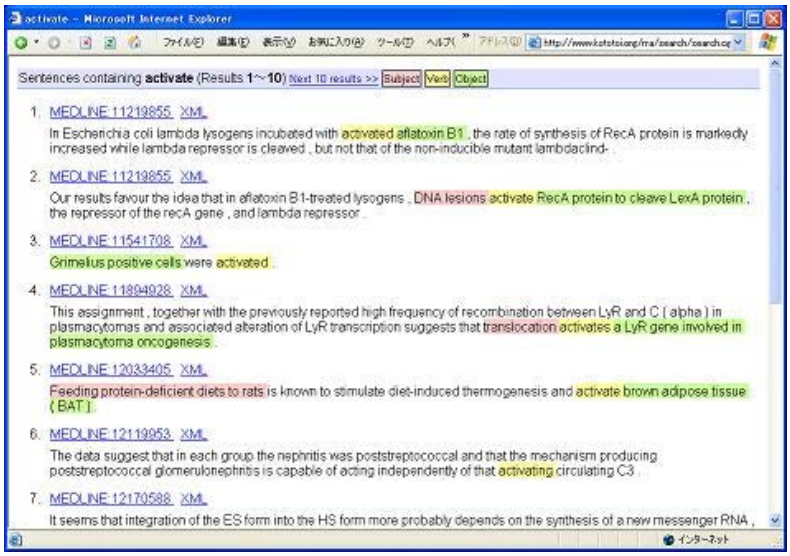



［図 3.8.4］ 統合システム構成

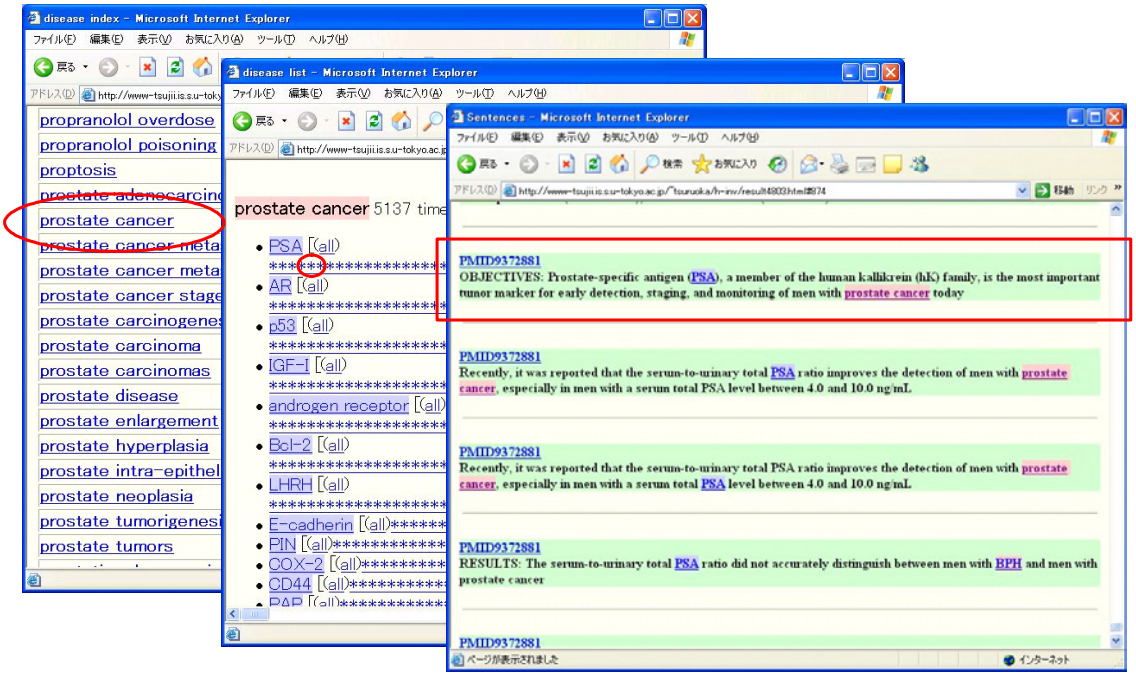
**検索・情報抽出**

GCL サーバには、xml で表現されている MEDLINE の全テキストが格納されている。各文における句構造、述語・項構造、固有表現に関する情報がインデックスされており、GCL クエリーによってユーザが指定したテキストを高速に検索することができる。たとえば、GCL を利用することで「動詞 active の意味上の主語と述語がたんぱく質である文」などを検索するようなことも可能である（図 3.8.5）。このような検索は単語のインデックスをベースとする通常の検索エンジンでは実現することは困難である。

GCL サーバがテキストに対するアノテーション情報を保持しているのに対し、DB サーバでは、テキストとは独立した情報を保持する役割を持つ。具体的には、エンティティに関する統計情報、ユーザが構築した知識ネットワークなどを保持する。たとえば、すでにテキストから抽出されている疾患と遺伝子の関係情報（図 3.8.6）などは、DB サーバーで保持される。




［図 3.8.5］クエリ”動詞:activate”に対する結果




［図 3.8.6］遺伝子・疾患の関係情報

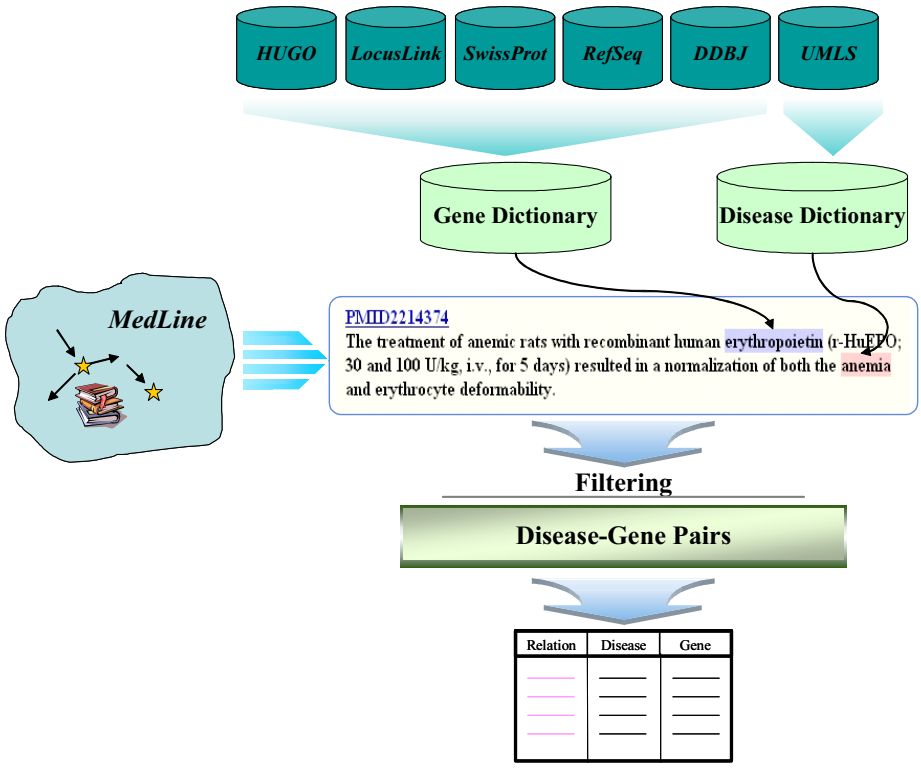
### ④ 疾患・遺伝子関係のテキストからの抽出

疾患と遺伝子の関係を抽出することは、医学・生物学分野におけるテキストマイニングの重要な課題の一つであり、そのような関係をテキストから発見するためのシステムが、ArrowSmith, BITOLA などをはじめとしてすでにいくつか実用化されている。既存の関係発見システムの多くは、関係情報の抽出を用語同士の共起情報に頼っているために、得られた結果に多くのノイズが含まれているという問題がある。BIONLP の分野では、機械学習とルールベースの手法などを組み合わせて、精度の高い情報抽出を行う手法も研究されているが、抽出結果がテキストの中で閉じていることが多く、その場合、抽出結果と生物学データベースとがリンクしていない。抽出結果が、医学生物学研究の

上で有用かつ再利用可能であるためには、テキスト中で認識されているエントリと、生物学データベースのエントリとの間に対応がついていることが不可欠である。

本研究では、生物学データベースから構築した辞書と機械学習手法を組み合わせることにより、生物学データベースと対応がついている遺伝子・疾患の関係情報を、高精度でテキストから抽出することを目的とする。

図 3.8.7 に、本アプローチの全体像を示す。




［図 3.8.7］ 遺伝子・疾患関係の抽出

**辞書の構築**

遺伝子辞書の構築には、5 つのパブリックデータベース（HUGO, LocusLink, SwissProt, RefSeq, DDBJ）を利用した。エントリの数は、34,959 エントリである。

疾患名辞書には、UMLS（2003AC Edition）を利用した。UMLS の Metathesaurus から、Semantic タイプとして、疾患や症状に関する TUI をもつエントリを抽出した。その結果、文字列の数として 431, 429 個、概念数として 159, 448 からなる疾患名辞書が構築された。

**コーパスアノテーション**

機械学習のトレーニングと評価のため、アノテーション付きコーパスを作成した。最初に MEDLINE から次のクエリーによって、疾患に関係するアブストラクトを抽出する。“Disease Category”[MeSH] AND (“Amino Acides, Peptides, and Proteins”[MeSH] OR “Genetic Structures”[MeSH])

得られた 1, 362, 285 アブストラクトに対して、遺伝子・疾患辞書と、最長一致法を利用することで、テキスト中の遺伝子名候補、病名候補の認識を行い、遺伝子と疾患の共

起の候補集合を作る。この中から 1000 個の共起をランダムサンプリングして、アノテーションを行う対象とした。アノテータは、遺伝子名候補、疾患名候補に対して、それらが、正しく遺伝子名（あるいは病名）として認識されているかどうかを２値でアノテーションを行う。さらに、遺伝子名・疾患名がともに正しい場合、その共起が遺伝子・疾患関係を表しているかどうかについても２値でアノテーションを行う。遺伝子が疾患に関係しているかどうかの判断については、以下の３種類の関係のいずれかが記述されているかどうかを基準とした。

- 病態生理学、疾患のメカニズム
- 遺伝子もしくはタンパク質の治療的効果
- 疾患リスク、診断、予後診断におけるマーカーとして役割

アノテーションの結果、遺伝子名と病名が両方とも正しく認識されている場合、94%の割合で、その共起が遺伝子と疾患の関係を表現していることが明らかになった。そのため、本タスクにおいては、フィルタリングの対象は、遺伝子・疾患名の認識のみとし、遺伝子と疾患の関係に関してはフィルタリングは行わない。

**機械学習によるフィルタリング**

用語認識のフィルタリングのための機械学習手法には、最大エントロピー法を利用した。判別のための特徴量としては、以下の情報を用いた。

- 遺伝子名認識
  - 遺伝子名候補自身、周辺 unigram/bigram、品詞、head、候補を argument としている述語、expanded form（候補が acronym の場合）
- 疾患名認識
  - 疾患名候補自身、周辺 unigram/bigram、品詞、head

**実験結果**

表 3.8.1 に疾患名、遺伝子名の認識精度を 10-fold の交差検定で評価した結果を示す。ここで、再現率ではなく相対再現率を評価指標としているのは、そもそも、アノテーションコーパスを作る際に、辞書マッチで候補の認識を行っているために、真の意味での再現率は計算できないからである。結果を見ると、遺伝子名、疾患名ともに、高い精度で機械学習による判定ができていることがわかる。

| | 適合率 | 相対再現率 |
|---|---|---|
| 遺伝子名 | 89.0 | 90.9 |
| 疾患名 | 90.0 | 96.6 |

**［表 3.8.1］ 遺伝子・疾患名の認識精度**

次に、遺伝子・疾患関係の抽出に関する精度を**表 3.8.2 に**示す。先に述べたように、抽出システムは、遺伝子名・疾患名のどちらともが正しいと判定された場合、無条件にそれらの間に関係があるとみなす。結果を見ると、適合率が 50%程度から 80%近くへと大きく向上していることがわかる。相対再現率は１割程度しか低下してない。

| | 適合率 | 相対再現率 |
|---|---|---|
| フィルタリングなし | 51.8% | 100.0% |
| フィルタリングあり | 78.5% | 87.1% |

［表 3.8.2］ 疾患・遺伝子関係抽出の精度

## ［関連発表文献］

(Chun et al., 2006) Chun, Hong-Woo, Yoshimasa Tsuruoka, Jing-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Jun'ichi Tsujii. (2006). **Extraction of Gene-Disease Relations from Medline using Domain Dictionaries and Machine Learning.** In the Proceedings of the Pacific Symposium on Biocomputing (PSB) 2006. (to appear)

### (2) 研究成果の今後期待される効果

統合サービスシステムは、開発された要素技術が実際のユーザの要求を満足させるシステム構築に有効であることを確認すると同時に、次に研究すべき技術の方向を見極める上で大きな役割を果たす。

本節の病疾患・遺伝子関係の発見援助を行うシステムは、産業総合研究所・ＪＢＩＲＣの五條堀グループとの共同で行ったものであり、医療・バイオ研究者が使用することで、ユーザの要求を取り入れたものにしていく予定である。とくに、データマイニングとテキスト処理、バイオオントロジーとのより有機的な結合、文脈を考慮した関係の認識など、新たな研究テーマが同定されている。この方向での研究は、生命科学のためのテキストマイングの研究として、大きく広がりつつある。

また、今回の統合サービスシステムでは、ユーザグループとの研究協力がもっとも進んでいる病疾患・遺伝子の問題を取り上げたが、テキストからの蛋白質相互作用の自動抽出、GENE オントロジーによる蛋白質の機能認定の２つシステムが、同じ Enju の文解析結果をもとに開発されている。

これらの異なったタスクのためめサービスシステムが、同じ言語処理の結果を基盤に構成できることは、情報要求時点以前に、かなりの部分の言語処理を大規模テキストに適用、それを知的な索引構造に反映させることで、効率的、かつ、高精度・知的な処理をサービス時点で可能にするという、本プロジェクトの当初の野心的な目標が現実的な目標であったことを示している。

## 4　研究参加者

### (1) 言語処理グループ

| 氏　名 | 所　属 | 役　職 | 研究項目 | 参加時期 |
|---|---|---|---|---|
| 辻井潤一 | 東京大学大学院<br>情報学環 | 教授 | 言語処理と統括 | 平成12年11月～<br>平成17年10月 |
| 西田豊明 | 京都大学大学院<br>情報学研究学科 | 教授 | 実世界エージェント技術 | 平成12年11月～<br>平成17年10月 |
| 宮尾祐介 | 東京大学大学院<br>情報学環 | 助手 | 文法フォーマリズム | 平成12年11月～<br>平成17年10月 |
| 鳥澤健太郎 | 北陸先端技術<br>大学院大学 | 助教授 | データベースの構築 | 平成12年11月～<br>平成17年10月 |
| 美馬秀樹 | 東京大学大学院<br>新領域創成科学 | 助手 | データベースの構築 | 平成13年 4月～<br>平成17年10月 |
| 建石由佳 | 東京大学大学院<br>情報理工学系研究科 | CREST研究員 | オントロジーの作成 | 平成13年 4月～<br>平成17年10月 |
| 鶴岡慶雅 | 東京大学大学院<br>情報理工学系研究科 | CREST研究員 | 機械学習 | 平成14年 4月～<br>平成17年10月 |
| 二宮崇 | 東京大学大学院<br>情報理工学系研究科 | CREST研究員 | 情報抽出 | 平成13年 4月～<br>平成17年10月 |
| Jin Dong Kim | 東京大学大学院<br>情報理工学系研究科 | CREST研究員 | 情報抽出 | 平成13年 2月～<br>平成15年 3月<br>平成16年 5月～<br>平成17年10月 |
| 大田朋子 | 東京大学大学院<br>情報理工学系研究科 | CREST研究員 | コーパスの作成 | 平成15年 4月～<br>平成17年10月 |
| 風間淳一 | 北陸先端技術<br>大学院大学 | 助手 | 情報抽出 | 平成12年11月～<br>平成17年10月 |
| 吉永直樹 | 北陸先端技術<br>大学院大学 | 学振研究員 | 文法フォーマリズム | 平成12年11月～<br>平成17年10月 |
| 薬師寺あかね | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 情報抽出 | 平成12年11月～<br>平成17年10月 |
| 狩野芳伸 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 文法学習 | 平成12年11月～<br>平成17年10月 |
| 荒木淳子 | 東京大学大学院<br>情報学環 | 博士課程 | 自動抄録 | 平成16年 4月～<br>平成17年10月 |
| 岡崎直観 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 情報抽出 | 平成16年 7月～<br>平成17年10月 |
| 増田勝也 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | テキストデータベース | 平成13年10月～<br>平成17年10月 |
| 松崎拓也 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 構文解析 | 平成14年 4月～<br>平成17年10月 |
| 全弘宇 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 情報抽出 | 平成16年 4月～<br>平成17年10月 |

| | | | | |
|---|---|---|---|---|
| 王悦 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | オントロジーの作成 | 平成17年 4月～<br>平成17年10月 |
| 綱川隆司 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 機械翻訳 | 平成14年10月～<br>平成17年10月 |
| 川崎敬昌 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | オントロジーの学習 | 平成14年10月～<br>平成16年 3月 |
| 吉田和弘 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 機械学習 | 平成14年10月～<br>平成17年10月 |
| 大内田賢太 | 東京大学大学院<br>情報理工学系研究科 | 博士課程 | 文法フォーマリズム | 平成14年10月～<br>平成17年10月 |
| 三輪誠 | 東京大学大学院<br>新領域創成科学研究科 | 博士課程 | WEBシステム | 平成14年 5月～<br>平成17年10月 |
| 原忠義 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 文法学習 | 平成13年10月～<br>平成17年10月 |
| 佐藤学 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 文法フォーマリズム | 平成15年10月～<br>平成17年10月 |
| 中西紘子 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 文法フォーマリズム | 平成15年10月～<br>平成17年10月 |
| 三浦研璽 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | オントロジー学習 | 平成15年10月～<br>平成17年10月 |
| 松林優一朗 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 情報抽出 | 平成16年 4月～<br>平成17年10月 |
| 新井元基 | 京都大学大学院<br>情報学研究科 | 修士課程 | 実世界のエージェント技術 | 平成16年 6月～<br>平成17年10月 |
| 大高雄介 | 京都大学大学院<br>情報学研究科 | 修士課程 | 実世界のエージェント技術 | 平成16年 6月～<br>平成17年10月 |
| 熊谷賢 | 京都大学大学院<br>情報学研究科 | 修士課程 | 実世界のエージェント技術 | 平成16年 6月～<br>平成17年10月 |
| 小関悠 | 京都大学大学院<br>情報学研究科 | 修士課程 | 実世界のエージェント技術 | 平成16年 6月～<br>平成17年10月 |
| 斎藤憲 | 京都大学大学院<br>情報学研究科 | 修士課程 | 実世界のエージェント技術 | 平成16年 6月～<br>平成17年10月 |
| 小嶋大起 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 文法フォーマリズム | 平成16年10月～<br>平成17年10月 |
| 竹内淳平 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 情報検索 | 平成16年10月～<br>平成17年10月 |
| 岡野原大輔 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 機械学習 | 平成16年10月～<br>平成17年10月 |
| 深町佳一朗 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 情報抽出 | 平成17年 4月～<br>平成17年10月 |
| 田谷滋規 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 情報抽出 | 平成17年 4月～<br>平成17年10月 |
| Philipp<br>Spanger | 東京大学大学院<br>情報理工学系研究科 | 研究生 | 文法フォーマリズム | 平成17年 4月～<br>平成17年10月 |

| | | | | |
|---|---|---|---|---|
| NguyenLuu Thuy Ngan | 東京大学大学院 情報理工学系研究科 | 研究生 | 機械翻訳 | 平成17年 4月～ 平成17年10月 |
| 干中華 | 東京大学大学院 情報理工学系研究科 | 研究生 | コーパスの作成 | 平成15年 6月～ 平成15年 9月 |
| 永野圭一郎 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 自動抄録 | 平成13年 4月～ 平成16年 3月 |
| 福林一平 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 文法フォーマリズム | 平成13年 4月～ 平成16年 3月 |
| 池田泰之 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 機械学習 | 平成15年10月～ 平成16年 3月 |
| 高野京子 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 生命科学の テキストベース作成 | 平成14年 5月～ 平成15年 6月 |
| 山田香己由 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 生命科学の テキストベース作成 | 平成14年 5月～ 平成15年 6月 |
| 若木裕美 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 自然言語処理ソフト | 平成14年 6月～ 平成15年 6月 |
| 窪田悠介 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 生命科学の テキストベース作成 | 平成14年 4月～ 平成15年 6月 |

### (2) 広域ソフトウェアグループ

| 氏　名 | 所　属 | 役　職 | 研究項目 | 参加時期 |
|---|---|---|---|---|
| 米澤明憲 | 東京大学大学院 情報理工学系研究科 | 教授 | 広域分散移動ソフト | 平成12年11月～ 平成17年10月 |
| 田浦健次朗 | 東京大学大学院 情報理工学系研究科 | 助教授 | 広域分散システム | 平成12年11月～ 平成17年10月 |
| 増原英彦 | 東京大学大学院 総合文化研究科 | 助教授 | 移動計算 | 平成12年11月～ 平成17年10月 |
| 高橋俊行 | 東京大学大学院 情報理工学系研究科 | CREST研究員 | 広域分散システムの 構築 | 平成13年 7月～ 平成17年10月 |
| 藤本浩史 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 広域分散システムの 構築 | 平成15年10月～ 平成17年 3月 |
| 遠藤敏夫 | 学振 | 特別研究員 | 広域分散システム | 平成12年11月～ 平成14年 3月 |
| 洪淳祥 | 東京大学大学院 情報理工学系研究科 | 修士課程 | WEBシステム | 平成12年11月～ 平成14年 3月 |
| 塩谷沢生 | 東京大学大学院 情報理工学系研究科 | 修士課程 | WEBシステム | 平成13年12月～ 平成15年 3月 |
| 遠藤侑介 | 東京大学大学院 情報理工学系研究科 | 修士課程 | WEBシステム | 平成17年 4月～ 平成17年10月 |
| 吉野寿宏 | 東京大学大学院 情報理工学系研究科 | 修士課程 | WEBシステム | 平成17年 4月～ 平成17年10月 |
| 山崎孝裕 | 東京大学大学院 情報理工学系研究科 | 修士課程 | 広域分散システムの 構築 | 平成17年 4月～ 平成17年10月 |

| | | | | |
|---|---|---|---|---|
| 沈垣甫 | 東京大学大学院<br>新領域創成科学研究科 | 修士課程 | 広域分散システムの構築 | 平成17年 4月～<br>平成17年10月 |

### (3) オントロジーグループ

| 氏　名 | 所　属 | 役　職 | 研究項目 | 参加時期 |
|---|---|---|---|---|
| 中川裕志 | 東京大学<br>情報基盤センター | 教授 | テキストからのオントロジー自動獲得 | 平成15年 4月～<br>平成17年10月 |
| 中田圭一 | ドイツ国際大学 | 教授 | オントロジー管理システム | 平成12年11月～<br>平成17年10月 |
| 田中久美子 | 東京大学<br>情報理工学系研究科 | 助教授 | 用例検索システムの設計 | 平成15年 4月～<br>平成17年10月 |
| 吉田稔 | 東京大学<br>情報基盤センター | 助手 | WEBページ構造解析学習アルゴリズム | 平成12年11月～<br>平成17年10月 |
| 清田陽司 | 東京大学<br>情報基盤センター | 助手 | 用例検索システムの実装 | 平成17年 4月～<br>平成17年10月 |
| 星野綾子 | 東京大学大学院<br>情報学府 | 修士課程 | WEBからの情報抽出 | 平成16年 9月～<br>平成17年10月 |
| 一井崇 | 東京大学大学院<br>情報学府 | 修士課程 | 用例検索システムの設計 | 平成16年 9月～<br>平成17年10月 |
| 藤本宏涼 | 東京大学大学院<br>情報学府 | 修士課程 | 用例検索システムの応用 | 平成16年 9月～<br>平成17年10月 |
| 小野真吾 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 人物オントロジー抽出の検討 | 平成16年 9月～<br>平成17年10月 |
| 佐々木寛子 | 東京大学大学院<br>情報理工学系研究科 | 修士課程 | 用例検索システムの応用 | 平成17年 4月～<br>平成17年10月 |
| 國安結 | 東京大学大学院<br>情報学府 | 修士課程 | 用例検索システムの応用 | 平成17年 4月～<br>平成17年10月 |
| 田枝覚 | 東京大学大学院<br>情報学府 | 非常勤職員 | 用例検索システムの実装 | 平成15年 6月～<br>平成16年 1月 |
| Klaus Voss | 東京大学<br>情報理工学系研究科 | CREST研究員 | オントロジー変換、システムの作成 | 平成13年 4月～<br>平成15年 3月 |
| 白山大地 | 東京大学<br>情報理工学系研究科 | 修士課程 | オントロジー工学 | 平成14年 4月～<br>平成15年 3月 |
| 岩越守孝 | 東京大学大学院<br>情報学府 | 修士課程 | 専門用語抽出 | 平成15年 7月～<br>平成15年11月 |
| 山本真人 | 東京大学大学院<br>情報学府 | 修士課程 | 用例検索システムの実装 | 平成15年 6月～<br>平成16年 3月 |
| 辻河亨 | 東京大学大学院<br>情報学府 | 修士課程 | 専門用語抽出 | 平成15年 6月～<br>平成16年 3月 |
| 河内崇 | 東京大学大学院<br>情報学府 | 修士課程 | 用例検索システムの実装 | 平成15年 6月～<br>平成17年 3月 |
| 森山聡 | 東京大学大学院<br>情報学府 | 修士課程 | 専門用語抽出 | 平成16年 4月～<br>平成17年 3月 |

# ５　成果発表等

## ５．１ 論文発表　（国内　14 件、海外　35 件）


1. Kenjiro Taura, Andrew Chien: A Heuristic Algorithm for Mapping Communicating Tasks on Heterogeneous Resources. In Proceedings of Heterogeneous Computing Workshop 2000, pages102-115, 2000.5
2. Takashi Ninomiya, Kentaro Torisawa and Jun'ichi Tsujii. An Agent-based Parallel HPSG Parser for Shared-memory Parallel Machines. Journal of Natural Language Processing. 8(1). 2001.
3. Hideki Mima, Sophia Ananiadou and Goran Nenadic. The ATRACT Workbench: An Automatic Term Recognition and Clustering of Terms. In the Text Speech and Dialogue (TSD 2001), Lecture Notes in Artificial Intelligence. 2166. pp. 126--133. Springer Verlag. 2001.
4. Kenji Kaneda, Kenjiro Taura, Akinori Yonezawa: Virtual Private Grid: A Command Shell for Utilizing Hundreds of Machines Efficiently. In Proceedings of 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2002), pages212-219, 2002
5. Naoki Yoshinaga and Yusuke Miyao. Grammar conversion from LTAG to HPSG. WEB-SLS: the European Student Journal on Language and Speech. 2002.
6. Jin-Dong Kim, Tomoko Ohta, Yuka Teteisi and Jun'ichi Tsujii. GENIA corpus - a semantically annotated corpus for bio-textmining. Bioinformatics. 19(suppl. 1). pp. i180-i182. Oxford University Press. 2003.
7. Naoki Yoshinaga, Yusuke Miyao, Kentaro Torisawa, Jun'ichi Tsujii. Parsing Comparison across Grammar Formalisms Using Strongly Equivalent Grammars. Journal of Traitement Automatique des Langues. 44(3). pp. 15--39. Association pour le Traitement Automatique des LAngues. 2003.
8. Toru Hisamitsu and Tsujii, Jun-ichi. Measuring Term Representativeness. In Pazienza, Maria Teresa (Eds.), Information Extraction in the Web Era. LNAI 2700. pp. 45-76. Springer-Verlag. 2003.
9. Minoru Yoshida, Kentaro Torisawa, Jun'ichi Tsujii. Extracting Attributes and Their Values from Web Pages. In A. Antonacopoulos, Jianying Hu (Eds.), Web Document Analysis: Challenges and Opportunities. pp. 179-200. World Scientific. 2003.
10. Hiroshi Nakagawa, Tatsunori Mori. Automatic Term Recognition based on Statistics of Compound Nouns and their Components, Terminology, Vol.9 No.2, pp. 201-219, 2003/4
11. Kenji Kaneda, Kenjiro Taura, Akinori Yonezawa: Virtual private grid: a command shell for utilizing hundreds of machines efficiently. Future Generation Computer Systems 19(4), pages 563-573, SanDiego, 2003.5
12. Kenjiro Taura, Toshio Endo, Kenji Kaneda, Akinori Yonezawa: Phoenix : a Parallel Programming Model for Accommodating Dynamically Joining/Leaving Resources. In Proceedings of ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP 2003), pages 216-229, 2003.6
13. Takashi Hoshino, Kenjiro Taura, Takashi Chikayama: An Adaptive File Distribution Algorithm for Wide Area Network. Proceedings of Workshop on Adaptive Grid Middleware,New Orleans,2003.9
14. Takashi Hoshino, Kenjiro Taura, Takashi Chikayama: An Adaptive File Distribution Algorithm for Wide Area Network. Spcial Issue of Journal of Parallel and Distributed Computing Practices,To appear ,2003
15. Minoru Yoshida, Kentaro Torisawa and Jun'ichi Tsujii, Extracting Attributes and Their Values from Web Pages,Chapter in a book titled Web Document Analysis: Challenges and Opportunities, A. Antonacopoulos and Jianying Hu, editors, Series in Machine Perception and Artificial Intelligence, World Scientific, pages 179-200, 2003/9

16. Yusuke Miyao, Takashi Ninomiya and Jun'ichi Tsujii. Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2004. LNAI3248. pp. 684-693. Springer-Verlag.
17. Yoshimasa Tsuruoka and Jun'ichi Tsujii. Improving the Performance of Dictionary-based Approaches in Protein Name Recognition. Journal of Biomedical Informatics. 37(6). pp. 461-470. Elsevier. 2004.
18. Yoshida, Minoru, Kentaro Torisawa, Jun'ichi Tsujii. Integrating Tables on the World Wide Web. Transactions of the Japanese Society for Artificial Intelligence. 19(6). pp. 548-560. The Japanese Society for Artificial Intelligence. 2004.
19. Takashi Masuyama, Hiroshi Nakagawa: "Two Step POS Selection for SVM based Text Categorization", IEICE Transaction of Special Issue on Information Processing Technology for Web Utilization, Vol. E87-D, pp.15-21, 2004/2
20. Toshio Endo, Kenji Kaneda, Kenjiro Taura, Akinori Yonezawa: High Performance LU Factorization for Non-Dedicated Clusters,IEEE/ACM Symposium on Cluster Computing and the Grid (CCGrid2004), Chicago, pp.678-685, 2004.4
21. Kenji Kaneda, Toshio Endo, Kenjiro Taura, Akinori Yonezawa: Routing and Resource Discovery in Phoenix Grid-Enabled Message Passing Library, IEEE/ACM Symposium on Cluster Computing and the Grid (CCGrid2004),Chicago,pp.670-677,2004.4
22. Kenjiro Taura: GXP : An interactive shell for the grid environment. In proceedings of Innovative Architecture for Future Generation HighPerformance Processors and Systems (IWIA2004) ,Hawaii,pages59-67,2004
23. Minoru Yoshida and Hiroshi Nakagawa: "Specification Retrieval - How to Find Attribute-Value Information on the Web", Natural Language Processing -IJCNLP 2004, Keh-Yih Su Jun’ichi Tsujii Jong-Hyeok Lee OiYee Kwong (Eds.), Lecture Notes in Computer Science, Springer Berlin Heidelberg Vol.3248,pp.338-347, 2005/10
24. Toshio Endo, Kenjiro Taura: Highly Latency Tolerant Gaussian Elimination, Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing（Grid2005）, Seattle, pp. 91-98, 2005.11
25. Hideo Saito, Kenjiro Taura, Takashi Chikayama: Collective Operations for Wide-Area Message Passing Systems Using Adaptive Spanning Trees, Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing （Grid2005）, Seattle, pp. 40-48, 2005.11
26. Yuuki Horita, Kenjiro Taura, Takashi Chikayama: A Scalable and Efficient Self-Organizing Failure Detector for Grid Applications, Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing （Grid2005）, Seattle, pp. 202-210 , 2005.11
27. Irena Spasic, Sophia Ananiadou, and Jun'ichi Tsujii. MaSTerClass: a case-based reasoning system for the classification of biomedical terms. Bioinformatics. 21(11). pp. 2749-2758. Oxford University Press. 2005.
28. Jun-chi Tsujii and Sophia Ananiadou. Thesaurus or logical onotology, which do we need for mining text? Language Resources and Evaluation. 39(1). pp. 77-90. Springer SBM. 2005.
29. Jun'ichi Kazama and Jun'ichi Tsujii. Maximum Entropy Models with Inequality Constraints: A case study on text categorization. Machine Learning Journal special issue on Learning in Speech and Language Technologies. 60(1-3). pp. 169-194. Springer SBM. 2005.
30. Tadayoshi Hara, Yusuke Miyao and Jun'ichi Tsujii. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In Robert Dale, Kam-Fai Wong, Jian Su, Oi Yee Kwong (Eds.), Natural Language Processing – IJCNLP 2005. LNC3651. Springer-Verlag. pp. 199-210. 2005.
31. Daisuke Okanohara and Jun'ichi Tsujii. Assigning Polarity Scores to Reviews Using Machine Learning Techniques. In Robert Dale, Kam-Fai Wong, Jian Su and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2005. LNCS3651. Springer-Verlag. pp. 314-325. 2005.

32. Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. Syntax Annotation for the GENIA corpus. In Robert Dale, Kam-Fai Wong, Jian Su and Oi Yee Kwong (Eds.), Natural Language Processing-IJCNLP2005. LNCS3651. Springer-Verlag. pp. 222--227. 2005.

33. Jin-Dong Kim and Jun'ichi Tsujii. Word Folding: Taking the Snapshot of Words Instead of the Whole. In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2004. LNAI 3248. Springer-Verlag. pp. 406-415. 2005.

34. Yoshimasa Tsuruoka and Jun'ichi Tsujii. Iterative CKY Parsing for Probabilistic Context-Free Grammars. In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2004. LNAI 3248. pp. 52-60. Springer-Verlag. 2005.

35. Takashi Ninomiya, Yusuke Miyao and Jun'ichi Tsujii. A Persistent Feature-Object Database for Intelligent Text Archive Systems. In Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee and Oi Yee Kwong (Eds.), Natural Language Processing - IJCNLP 2004. LNAI3248. pp. 197--205. Springer-Verlag. 2005.

36. Yoshio Nakao. A Method for Related-passage Extraction based on Thematic Hierarchy. IPSJ Transactions on Databases. 42(SIG 10 (TOD 11)). pp. 39--53. 2001.

37. 中川裕志,森辰則,湯本紘彰, 出現頻度と連接頻度に基づく専門用語抽出",自然言語処理,Vol.10 No.1, pp. 27 - 45, 2003/1

38. 増田 英孝,塚本 修一,安富 大輔,中川 裕志, HTML の表形式データの構造認識と携帯端末表示への応用, 情報処理学会論文誌：データベース SIG12 (TOD19), pp.23-32, 2003/10

39. 小峰 恒,山田 剛一,絹川 博之,中川 裕志, 文書頻度と節長を利用した図書概要縮約方式", NII Journal Vol.8, pp.23-34, 2004/2

40. 風間 淳一, 宮尾 祐介, 辻井 潤一. 教師なし隠れマルコフモデルを利用した最大エントロピータグ付けモデル. 自然言語処理. 11(4). pp. 3-23. 2004.

41. 山田雅信, 田浦健次朗, 近山隆, 高橋俊行: インクリメンタルPageRankによる重要Webページの効率的な収集戦略. 先進的計算基盤システムシンポジウム (SACSIS2004), IPSJ Symposium Series, Vol.2004, No.6, pp.103-110, 2004.5

42. Hideo Saito, Kenjiro Taura, Takashi Chikayama: Expedite: An Operating System Extension to Support Low-Latency Communication in Non-Dedicated Clusters. 先進的計算基盤システムシンポジウム (SACSIS2004), IPSJ Symposium Series, Vol.2004, No.6, pp.443-450, 2004.5

43. 安藤雅享, 田浦健次朗, 近山隆: Grid 環境での並列ジョブ投入を支援するシェル. 先進的計算基盤システムシンポジウム (SACSIS2004), IPSJ Symposium Series, Vol.2004, No.6, pp.225-232, 2004.5

44. 山田雅信, 田浦健次朗, 近山隆, 高橋俊行: インクリメンタルPageRankによる重要Webページの効率的な収集戦略. In IPSJ Transactions on Advanced Computing Systems, Vol.45, No.SIG11(ACS7), pp.465-473,2004.10

45. Hideo Saito, Kenjiro Taura, Takashi Chikayama: Expedite: An Operating System Extension to Support Low-Latency Communication in Non-Dedicated Clusters. In IPSJ Transactions on Advanced Computing Systems, Vol.45, No.SIG11(ACS7), pp.229-237,2004.10

46. Minoru Yoshida, Kentaro Torisawa, and Jun'ichi Tsujii. Integrating Tables on the World Wide Web, 人工知能学会論文誌(Transactions of the Japanese Society for Artificial Intelligence). 19(6). pp. 548-560, 2004/10

47. 堀田勇樹, 田浦健次朗, 近山隆：耐故障並列計算を支援する自律的な故障検知機構. 先進的計算基盤システムシンポジウム (SACSIS2005),IPSJ Symposium Series, Vol.2005, No.5, pp.311-319,2005.5

48. 堀田勇樹, 田浦健次朗, 近山隆：耐故障並列計算を支援する自律的な故障検知機構. In IPSJ Transactions on Advanced Computing Systems, vol.46, No.SIG12(ACS11), pp.236-244, 2005

49. 中川裕志: "Kiwi:多言語用例検索システム" 漢字文献情報処理研究 6 号, pp.116-123, 2005/10

## ５．２ 口頭発表（国際学会発表及び主要な国内学会発表）

### (1) 招待、口頭講演 （国内 49 件、海外 78 件）

1. Naoki Yoshinaga, Yusuke Miyao, Kentaro Torisawa and Jun'ichi Tsujii. Resource sharing among HPSG and LTAG communities by a method of grammar conversion from FB-LTAG to HPSG. In the Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education. pp. 39--46. Morgan Kaufman Publishers. 2001.
2. Yoshio Nakao. How small a distinction among summaries can an IR-based evaluation method identify? In the Proceedings of NAACL 2001 Workshop on Automatic Summarization (WAS2001). pp. 69--78. 2001.
3. Kenji Nishida, Kentaro Torisawa and Jun'ichi Tsujii. Compiling an HPSG-based grammar into more than one CFG. In the Proceedings of PACLING 2001. pp. 199--206. 2001.
4. Naoki Yoshinaga, Yusuke Miyao, Kentaro Torisawa and Jun'ichi Tsujii. Efficient LTAG parsing using HPSG parsers. In the Proceedings of Pacific Association for Computational Linguistics (PACLING 2001). pp. 342--351. 2001.
5. Hideki Mima, Sophia Ananiadou and Goran Nenadic. Improving Knowledge Acquisition Through Automatic Term Recognition. In the Proceedings of Panhellenic Conference on Human Computer Interaction (PC-HCI 2001). pp. 177-182. 2001.
6. Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Hideki Mima and Jun'ichi Tsujii. Ontology Based Corpus Annotation and Tools. In the Proceedings of the 12th Genome Informatics 2001. pp. 469--470. 2001.
7. Yoshio Nakao. How small a distinction among summaries can the evaluation method identify? In the Proceedings of the NTCIR-2 workshop. pp. 341--348. 2001.
8. Minoru Yoshida, Kentaro Torisawa and Jun'ichi Tsujii. Extracting ontologies from World Wide Web via HTML tables. In the Proceedings of the Pacific Association for Computational Linguistics (PACLING 2001). pp. 332-341. 2001.
9. Jun'ichi Kazama, Yusuke Miyao and Jun'ichi Tsujii. A Maximum Entropy Tagger with Unsupervised Hidden Markov Models. In the Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium. pp. 333--340. 2001.
10. Minoru Yoshida, Kentaro Torisawa and Jun'ichi Tsujii. A method to integrate tables of the World Wide Web. In the Proceedings of the first International Workshop on Web Document Analysis (WDA 2001). pp. 31-34. 2001.
11. Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. XML-Based Linguistic Annotation of Corpus. In the Proceedings of the first NLP and XML Workshop held at NLPRS 2001. pp. 47--53. 2001.
12. Naoki Yoshinaga and Yusuke Miyao. Grammar conversion from LTAG to HPSG. In the Proceedings of the sixth ESSLLI Student Session. pp. 309--324. 2001.
13. Tomoko Ohta, Yuka Tateisi and Jun'ichi Tsujii. Tools for Ontology-based Corpus Annotation. In the Proceedings of the sixth Pacific Symposium on Biocomputing (PSB 2001). pp. 112. 2001.
14. Akane Yakushiji, Yuka Tateisi, Yusuke Miyao and Jun'ichi Tsujii. Event extraction from biomedical papers using a full parser. In the Proceedings of the sixth Pacific Symposium on Biocomputing (PSB 2001). pp. 408-419. 2001.
15. Takaki Makino, Kazuyuki Aihara and Jun'ichi Tsujii. Towards Sentence Understanding: Phase Arbitration in Temporal-Coding Memory Mechanism. In the The Second Workshop on Natural Language Processing and Neural Networks (NLPNN 2001). pp. 46--52. 2001.
16. Yusuke Miyao and Jun'ichi Tsujii. Maximum Entropy Estimation for Feature Forests. In the Proceedings of Human Language Technology Conference (HLT 2002). 2002.
17. Hideki Mima, Sohia Ananiadou, Goran Nenadic and Jun'ichi Tsujii. XML Tag Information Management System: A Workbench for Ontology-based Knowledge Acquisition and Integration. In the Proceedings of Human Language Technology Conference (HLT 2002). 2002.

18. Hideki Mima, Sophia Ananiadou, Goran Nenadic and Jun'ichi Tsujii. TIMS - A Workbench for Ontology-based Knowledge Acquisition and Integration. In the Proceedings of Natural Language Processing in Biomedical Applications (NLPBA 2002). 2002.

19. Takashi Ninomiya, Yusuke Miyao and Jun'ichi Tsujii. Lenient Default Unification for Robust Processing within Unification Based Grammar Formalisms. In the Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). pp. 744--750. 2002.

20. Takashi Ninomiya, Takaki Makino and Jun'ichi Tsujii. An Indexing Scheme for Typed Feature Structures. In the Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). pp. 1248-1252. 2002.

21. Junko Araki, Takashi Ninomiya, Takaki Makino and Jun'ichi Tsujii. Action Vectors for Interpreting Route Descriptions. In the Proceedings of the AAAI-02 Workshop on Spatial and Temporal Reasoning. 2002.

22. Minoru Yoshida. Extracting Attributes and Their Values from Web Pages. In the Proceedings of the ACL 2002 Student Research Workshop. pp. 72--77. 2002.

23. Tomoko Ohta, Yuka Tateisi, Hideki Mima and Jun'ichi Tsujii. GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. In the Proceedings of the Human Language Technology Conference (HLT 2002). 2002.

24. Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta and Jun'ichi Tsujii. Tuning Support Vector Machines for Biomedical Named Entity Recognition. In the Proceedings of the Natural Language Processing in the Biomedical Domain (ACL 2002). pp. 1-8. 2002.

25. Jin-Dong Kim and Jun'ichi Tsujii. Corpus-Based Approach to Biological Entity Recognition. In the Proceedings of the Second Meeting of the Special Interest Group on Text Data Mining of ISMB 2002. 2002.

26. Tadayoshi Hara, Yusuke Miyao and Jun'ichi Tsujii. Clustering for obtaining syntactic classes of words from automatically extracted LTAG grammars. In the Proceedings of the sixth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6). pp. 227-233. 2002.

27. Junko Araki, Takashi Ninomiya, Takaki Makino and Jun'ichi Tsujii. Two Perspective systems Using a Route as a Reference Object. In the Proceedings of the sixth World Multiconference on Systemics, Cybernetics and Informatics (SCI 2002). 2002.

28. Naoki Yoshinaga, Yusuke Miyao and Jun'ichi Tsujii. A formal proof of strong equivalence for a grammar conversion from LTAG to HPSG-style. In the Proceedings of the sixth International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6). pp. 187--192. 2002.

29. M. Yoshida, Extracting Attributes and Their Values from Web PagesIn Proceedings of the ACL-02 Student Research Workshop, pages 72-77, 2002/7

30. Hiroshi Nakagawa and Tataunori Mori, A Simple but Powerful automatic Term Extraction Method, Computerm2: 2nd International Workshop on Computational Terminology, COLING-2002 WORKSHOP, pp. 29-35,Taipei, 2002/8

31. Takashi Masuyama and Hiroshi Nakagawa,Applying Cascaded Feature Selection to SVM Text Categorization", 3rd International Workshop on Natural Language and Information Systems, pp. 241-245, Aix-en-Provence, France, 2002/9

32. Yoshiaki Kawasaki, Jun'ichi Kazama and Jun'ichi Tsujii. Extracting Biomedical Ontology from Textbooks and Article Abstracts. In the Proceedings of the SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics. pp. 44-50. 2003.

33. Akane Yakushiji, Yuka Tateisi, Yusuke Miyao, Naoki Yoshinaga and Jun'ichi Tsujii. A Debug Tool for Practical Grammar Development. In the Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics. pp. 173--176. 2003.

34. Jun-ichi Tsujii. New Perspectives of Linguistic Study. In the ISCA and IEEE workshop on Spontaneous Speech Processing and Recognition. pp. 151-157. 2003.

35. Katsuya Masuda, Takashi Ninomiya, Yusuke Miyao, Tomoko Ohta and Jun'ichi Tsujii. A Robust Retrieval Engine for Proximal and Structural Search. In the Proceedings of HLT-NAACL 2003 Short papers. pp. 58--60. 2003.

36. Yusuke Miyao, Takashi Ninomiya and Jun'ichi Tsujii. Lexicalized Grammar Acquisition. In the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) companion volume. pp. 127--130. 2003.

37. Jun'ichi Kazama and Jun'ichi Tsujii. Evaluation and Extension of Maximum Entropy Models with Inequality Constraints. In the Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003). pp. 137-144. 2003.

38. Yoshimasa Tsuruoka and Jun'ichi Tsujii. Probabilistic Term Variant Generator for Biomedical Terms. In the Proceedings of the 26th Annual International ACM SIGIR Conference. pp. 167--173. 2003.

39. Naoki Yoshinaga, Kentaro Torisawa and Jun'ichi Tsujii. Comparison between CFG filtering techniques for LTAG and HPSG. In the Proceedings of the 41st ACL companion volume. pp. 185--188. 2003.

40. Jin-Dong Kim, Hae-Chang Rim and Jun'ichi Tsujii. Self-Organizing Markov Models and Their Application to Part-of-Speech Tagging. In the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. pp. 296-302. 2003.

41. Katsuya Masuda. A Ranking Model of Proximal and Structural Text Retrieval Based on Region Algebra. In the Proceedings of the ACL 2003 Student Research Workshop. pp. 50--57. 2003.

42. Yoshimasa Tsuruoka and Jun'ichi Tsujii. Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In the Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine. pp. 41-48. 2003.

43. Zhonghua Yu, Yoshimasa Tsuruoka and Jun'ichi Tsujii. Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense Per Discourse Hypothesis. In the Proceedings of the SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics. pp. 57-62. 2003.

44. Yusuke Miyao and Jun'ichi Tsujii. A model of syntactic disambiguation based on lexicalized grammars. In the Proceedings of the Seventh Conference on Natural Language Learning (CoNLL) at HLT-NAACL 2003. pp. 1--8. 2003.

45. Takuya Matsuzaki, Yusuke Miyao and Jun'ichi Tsujii. An Efficient Clustering Algorithm for Class-based Language Models. In the Proceedings of the Seventh Conference on Natural Language Learning (CoNLL) at HLT-NAACL 2003. pp. 119--126. 2003.

46. Yoshimasa Tsuruoka and Jun'ichi Tsujii. Training a Naive Bayes Classifier via the EM Algorithm with a Class Distribution Constraint. In the Proceedings of the Seventh Conference on Natural Language Learning (CoNLL) at HLT-NAACL 2003. pp. 127--134. 2003.

47. Yusuke Miyao, Takashi Ninomiya and Jun'ichi Tsujii. Probabilistic modeling of argument structures including non-local dependencies. In the the Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP) 2003. pp. 285--291. 2003.

48. Takashi Masuyama and Hiroshi Nakagawa, Cascaded Feature Selection in SVM Text Categorization, Lecture Note in Computer Science: the proceedings of 4th International Coference on Intelligent Text Processing and Computational Linguistics CICLing-2003, pp.588-591, Mexico City, 2003/2

49. Kumiko Tanaka-Ishii, Michiko Abe, and Hiroshi Nakagawa., Categorization of movies using comments, Proceedings of PACLING'03 (Pacific Association for Computational LINGuistics), pp.221-229, Halifax, Nova Scotia, Canada, 2003/8 (Best Paper Award of PACLING'03)

50. Hiroko Nakanishi, Yusuke Miyao and Jun'ichi Tsujii. Using Inverse Lexical Rules to Acquire a Wide-coverage Lexicalized Grammar. In the IJCNLP 2004 Workshop on Beyond Shallow Analyses. 2004.

51. Yuka Tateisi, Ohta, Tomoko and Tsujii, Jun-ichi. Annotation of Predicate-argument Structure of Molecular Biology Text. In the IJCNLP-04 workshop on Beyond Shallow Analyses. 2004.

52. Jun-ichi Tsujii. Thesaurus or logical ontology, which do we need for mining text? (keynote speech). In the Proc. of Language resources and evaluation conference (LREC 2004). Vol.III. pp. pp IX-XVI. 2004.

53. Yuka Tateisi and Jun'ichi Tsujii. Part-of-Speech Annotation of Biology Research Abstracts. In the Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004). IV. pp. 1267-1270. 2004.

54. Yusuke Miyao and Jun'ichi Tsujii. Deep Linguistic Analysis for the Accurate Identification of Predicate-Argument Relations. In the Proceedings of COLING 2004. pp. 1392-1397. 2004.

55. Yoshimasa Tsuruoka, Yusuke Miyao and Jun'ichi Tsujii. Towards efficient probabilistic HPSG parsing: integrating semantic and syntactic preference to guide the parsing. In the Proceedings of IJCNLP-04 Workshop: Beyond shallow analyses - Formalisms and statistical modeling for deep analyses. 2004.

56. Naoki Yoshinaga and Jun'ichi Tsujii. Generalizing Subcategorization Frames Acquired from Corpora Using Lexicalized Grammars. In the Proceedings of TAG+7. pp. 104--110. 2004.

57. Kenta Oouchida, Naoki Yoshinaga and Jun'ichi Tsujii. Context-free Approximation of LTAG towards CFG Filtering. In the Proceedings of TAG+7. pp. 171--177. 2004.

58. Jin-Dong Kim and Jun'ichi Tsujii. Word Folding: Taking the Snapshot of Words Instead of the Whole. In the Proceedings of the First International Joint Conference on Natural Language Processing. pp. 61--68. 2004.

59. Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi and Nigel Collier. Introduction to the Bio-Entity Recognition Task at JNLPBA. In the Proceedings of the International Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-04). pp. 70--75. 2004.

60. Hiroko Nakanishi, Yusuke Miyao and Jun'ichi Tsujii. An Empirical Investigation of the Effect of Lexical Rules on Parsing with a Treebank Grammar. In the Proceedings of the third TLT2004. pp. 103--114. 2004.

61. Naoki Yoshinaga. Improving the Accuracy of Subcategorizations Acquired from Corpora. In the The 42nd ACL Student Session. pp. 43--48. 2004.

62. Akane Yakushiji, Yuka Tateisi, Yusuke Miyao and Jun'ichi Tsujii. Finding Anchor Verbs for Biomedical IE Using Predicate-Argument Structures. In the the Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics. pp. 157--160. 2004.

63. Minoru Yoshida and Hiroshi Nakagawa. Specification Retrieval -- How to Find Attribute-Value Information on the Web?, Proceedings of IJCNLP (International Joint Conference of Natural Language Processing) 2004, pp.520-527, Hainan-Island, China, March 2004/3

64. Koichi Yamada, Hisashi Komine, Hiroshi Kinukawa and Hiroshi Nakagawa, Abstract of Abstract: A New Summarizing Method based on Document Frequency and Clause Length, SCI2004(The 8th World Multi-Conference on Systemics, Cybernetics and Informatics), Volume XIV, pp.56-61, Orland,U.S.A., 2004/7

65. Takeshi Masuyama, Satoshi Sekine and Hiroshi Nakagawa, Automatic Construction of Japanese KATAKANA Variant List from Large Corpus, COLING2004(Proceedings of the 20th International Conference on Computational Linguistics), pp. 1214-1219, Geneva, Switzerland, 2004/8

66. Hiroko Nakanishi, Yusuke Miyao and Jun'ichi Tsujii. Probabilistic models for disambiguation of an HPSG-based chart generator. In the Proc. of IWPT 2005. 2005.

67. Takashi Ninomiya, Yoshimasa Tsuruoka, Yusuke Miyao and Jun'ichi Tsujii. Efficacy of Beam Thresholding, Unification Filtering and Hybrid Parsing in Probabilistic HPSG Parsing. In the Proc. of IWPT 2005. 2005.

68. Yusuke Miyao and Jun'ichi Tsujii. Probabilistic disambiguation models for wide-coverage HPSG parsing. In the Proceedings of ACL 2005. pp. 83-90. 2005.

69. Yoshimasa Tsuruoka and Jun'ichi Tsujii. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. In the Proceedings of HLT/EMNLP 2005. pp. 467-474. 2005.

70. Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou and Jun'ichi Tsujii. Developing a Robust Part-of-Speech Tagger for Biomedical Text. Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, pp. 382-392. 2005.

71. Takuya Matsuzaki, Yusuke Miyao and Jun'ichi Tsujii. Probabilistic CFG with Latent Annotations. In the Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. pp. 75-82. 2005.

72. Yoshimasa Tsuruoka and Jun'ichi Tsujii. Chunk Parsing Revisited. In the Proceedings of the 9th International Workshop on Parsing Technologies (IWPT 2005). pp. 133-140. 2005.

73. Yoshimasa Tsuruoka, Sophia Ananiadou and Jun'ichi Tsujii. A Machine Learning Approach to Acronym Generation. In the Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. pp. 25-31. 2005.

74. Akane Yakushiji, Yusuke Miyao, Yuka Tateisi and Jun'ichi Tsujii. Biomedical Information Extraction with Predicate-Argument Structure Patterns. In the the Proceedings of the First International Symposium on Semantic Mining in Biomedicine. pp. 60--69. 2005.

75. Kumiko Tanaka-Ishii and Hiroshi Nakagawa, A Multilingual Usage Consultation Tool based on Internet Searching ---More than search engine, Less than QA, The 14th International World Wide Web Conference (WWW2005), Chiba, Japan, pp.363-371. 2005/5 (Best Presentation Awards of WWW2005)

76. Ayako Hoshino and Hiroshi Nakagawa, A real-time multiple-choice question generation for language testing -- a preliminary study--, 43rd ACL2005 Second Workshop on Building Educational Applications Using Natural Language Processing, pp.17-20. Ann Arbor, 2005/7

77. Takeshi Masuyama and Hiroshi Nakagawa, Web-based Acquisition of Japanese Katakana Variants, SIGIR2005 (The 28th Annual International ACM SIGIR Conference) . pp.338-344, Salvador, Brazil, 2005/8

78. Minoru Yoshida, Hiroshi Nakagawa, "Automatic Term Extraction based on Perplexity of Compound Words", IJCNLP'05 (The 2'nd International Joint Conference of Natural Language Processing, pp.269--279) Juje, Korea, 2005/10

79. 永野 圭一郎, 辻井 潤一, 鳥澤 健太郎. 談話文からの命題－様相の抽出システム. 言語処理学会第 7 回年次大会発表論文集. pp. 38--41. 2001.

80. 吉永 直樹, 宮尾 祐介, 鳥澤 健太郎, 辻井 潤一. LTAG 文法による HPSG パーザを用いた構文解析. 言語処理学会第 7 回年次大会発表論文集. pp. 277--280. 2001.

81. 新福 哲, 風間 淳一, 宮尾 祐介, 辻井 潤一. 多様な情報源を用いた隠れマルコフモデルによる医学/生物学論文の専門用語認識器. 言語処理学会第 8 回年次大会発表論文集. pp. 21--24. 2002.

82. 増田 勝也, 二宮 崇, 辻井潤一. GCL 問い合わせ代数による構造化テキストのランク検索. 言語処理学会第 8 回年次大会発表論文集. pp. 459--462. 2002.

83. 荒木 淳子, 二宮 崇, 牧野 貴樹, 辻井潤一. 経路の方向軸が説明する道順指示の曖昧性. 第 16 回人工知能学会全国大会. 2002.

84. 小峰恒,絹川博之,中川裕志,単語の文書頻度と文の長さを利用した抄録縮約方式, 情報処理学会 NL 研究会 NL-149, pp.73-80, 2002/5

85. 阿部倫子,田中久美子,中川裕志,コメントを用いた映画の分類, 情報処理学会 NL 研究会 NL-150, pp.105-110, 2002/7

86. 塚本修一,増田英孝,中川裕志, HTML の表形式データの変換と携帯端末表示への応用, 情報処理学会自然言語処理研究会 NL-151, pp.35-42, 2002/9

87. 山本真人,田中久美子,中川裕志, web 検索に基づく多言語動的 KWIC, 情報処理学会 NL 研究会 NL-152, pp.115-122, 2002/11

88. 塚本修一,増田英孝,中川裕志,"携帯端末表示を目指した HTML の表形式データの構造認識と変換", 情報学シンポジウム,pp.5-8,情報処理学会情報学基礎研究会主催, 2003/1

89. 山本真人, 田中久美子, 中川裕志, 検索エンジンに基づく多言語用例指南ツール:Kiwi",言語処理学会第 9 回大会発表番号 A1-5, 2003.4,（優秀発表賞）

90. 吉田 和弘, 鶴岡 慶雅, 宮尾祐介, 辻井潤一. 確率的言語モデルにおけるパラメータの確率分布を推定する手法とその応用. 言語処理学会第 9 回年次大会論文集. pp. 250--253. 2003.

91. 坂尾 要祐, 宮尾 祐介, 辻井潤一. 構文解析における非局所的な評価値の構成的計算. 言語処理学会第 9 回年次大会論文集. pp. 549--552. 2003.

92. 大内田 賢太, 吉永 直樹, 二宮 崇, 宮尾 祐介, 辻井 潤一. 語彙化文法における語彙項目の構造的特徴に基づく自動分類. 言語処理学会第 9 回年次大会論文集. pp. 553--556. 2003.

93. 薬師寺 あかね, 建石 由佳, 宮尾 祐介, 吉永 直樹, 辻井 潤一. 実用的な文法を開発するためのデバッグツール. 情報処理学会研究報告 NL-155. pp. 19--24. 2003.

94. 綱川 隆司, 二宮 崇, 宮尾 祐介, 辻井 潤一. 情報検索のためのトピック適応型単語クラスタリング. 言語処理学会第 10 回年次大会発表論文集. pp. 21-24. 2004.

95. 三浦 研爾, 鶴岡 慶雅, 辻井 潤一. 意味クラスタリングを用いた Web 文書からの概念階層の取得. 言語処理学会第 10 回年次大会発表論文集. pp. 333-336. 2004.

96. 大内田 賢太, 吉永 直樹, 辻井 潤一. CFG フィルタリングを用いた高効率な LTAG パーザの構築. 言語処理学会第 10 回年次大会発表論文集. pp. 361--364. 2004.

97. 増田 勝也, 二宮 崇, 宮尾 祐介, 辻井 潤一. 領域代数を用いた構造化テキスト検索の頑健でスケーラブルなモデル. 言語処理学会第 10 回年次大会発表論文集. pp. 377--380. 2004.

98. 辻河亨, 吉田稔, 中川裕志, 語彙空間の構造に基づく専門用語抽出情報処理学会 NL 研究会 159, pp.155-162, 2004/1

99. 上野友司, 森辰則, 木戸冬子, 中川裕志,係り受けの 2 部グラフと共起関係を利用した同義表現抽出": 情報処理学会 NL 研究会 159, pp.169-176, 2004/1

100. 吉田稔, 中川裕志, WWWからの属性・属性値情報ページの検索, 言語処理学会第 10 回年次大会発表論文集, pp. 620—623, 2004/3

101. 辻河亨, 吉田稔, 中川裕志, 統計的およびグラフ的素性を用いた専門用語抽出言語処理学会第 10 回年次大会発表論文集, pp. 33—36, 2004/3

102. 上野友司, 森辰則, 木戸冬子, 中川裕志,"係り受けの２部グラフと共起関係を利用した同義表現抽出", 言語処理学会 第 10 回年次大会, A1-1, 2004/3

103. 辻河亨, 吉田稔, 中川裕志, 統計的およびグラフ的素性を用いた専門用語抽出", 言語処理学会 第 10 回年次大会, A1-6, 2004/3

104. 小峰恒, 山田剛一, 絹川博之, 中川裕志, 重要節抽出型要約における読みやすさ向上方式の検討, 言語処理学会 第 10 回年次大会, C2-5, 2004/3

105. 山本真人, 田中久美子, 中川裕志, 多言語用例指南ツール:Kiwi の実験的評価, 言語処理学会第 10 回年次大会, D5-1, 2004/3

106. 吉田稔, 中川裕志, WWWからの属性・属性値情報ページの検索", 言語処理学会 第 10 回年次大会, D5-3, 2004/3

107. 森山聡,辻河亨,吉田稔,中川裕志, 専門用語抽出方法のテストコレクション依存性情報処理学会研究報告 「自然言語処理」No.161, 2004/5

108. 田浦健次朗: Grid Explorer: A Tool for Discovering, selecting, and using Distributed Resources Efficiently. 2004 年並列/分散/協調処理に関する「青森」サマー・ワークショップ(SWoPP 青森 2004), 情報処理学会研究報告 2004-HPC-99, Vol.2004, No.81, pp.235-240, 2004.8

109. 金田憲二, 田浦健次朗, 米沢明憲: 接続を動的に制御するメッセージパッシングシステム. 2004 年並列/分散/協調処理に関する「青森」サマー・ワークショップ(SWoPP 青森 2004), 情報処理学会研究報告 2004-HPC-99, Vol.2004, No.81, pp.241-246, 2004.8

110. 鴨志田良和, 田浦健次朗, 近山隆: 論理式の充足可能性問題の並列化における Clause 共有の効果について. 2004 年並列/分散/協調処理に関する「青森」サマー・ワークショップ(SWoPP 青森 2004), 電子情報通信学会技術研究報告, Vol.104, No.240, pp.25-30, 2004.8

111. 早津政和, 田浦健次朗, 近山隆: 実行時依存解析に基づく半自動並列化の効率的実装. 2004 年並列/分散/協調処理に関する「青森」サマー・ワークショップ(SWoPP 青森 2004), 情報処理学会論文誌: プログラミング, Vol.46, No.SIG 1(PRO 24),2004.8

112. 斎藤秀雄, 田浦健次朗, 近山隆: Phoenix プログラミングモデル用の集団通信. 2004 年並列/分散/協調処理に関する「青森」サマー・ワークショップ(SWoPP 青森 2004), 情報処理学会研究報告 2004-HPC-99, Vol.2004, No.81, pp.223-228, 2004.8

113. 堀田勇樹, 田浦健次朗, 近山隆: Phoenix プログラミングモデルにおける故障検知ライブラリ. 2004 年並列/分散/協調処理に関する「青森」サマー・ワークショップ(SWoPP 青森 2004), 情報処理学会研究報告 2004-HPC-99, Vol.2004, No.81, pp.229-234. 2004.8

114. 大田朋子, 建石由佳, 金進東, 薬師寺あかね, 辻井潤一. 生命科学分野のタグ付きコーパス : GENIA コーパスの設計と作成. 言語処理学会第 11 回年次大会発表論文集. pp. 506--509 2005.

115. 薬師寺あかね, 宮尾祐介, 建石由佳, 辻井潤一. 述語項構造パターンを用いた医学・生物学分野情報抽出. 言語処理学会第 11 回年次大会発表論文集. pp. 93--96. 2005.

116. 竹内淳平, 辻井潤一. 係り受け関係と言い換え関係を用いた柔軟な日本語検索. 言語処理学会第 11 回年次大会発表論文集. pp. 568--571. 2005.

117. 岡野原大輔, 辻井潤一. 評価文に対する二極指標の自動付与. 言語処理学会第 11 回年次大会発表論文集. pp. 664-667. 2005.

118. 森山聡,吉田稔, 中川裕志, "複合語パープレキシティに基づく重要語抽出法の研究", 言語処理学会 第 11 回年次大会, B2-8, 2005/3

119. 増山毅司, 中川裕志, "Web データを利用したカタカナ異表記の自動獲得", 言語処理学会 第 11 回年次大会, B2-9, 2005/3

120. 藤本宏凉,吉田稔, 中川裕志, "ローカルコーパスからのテキストマイニングツール : PortableKiwi", 言語処理学会 第 11 回年次大会, C1-8, 2005/3

121. 鴨志田良和,田浦健次朗,近山隆 : 論理式の充足可能性問題における変数の依存関係に基づく効率的な変数決定順序.2005 年並列/分散/協調処理に関する「武雄」サマー・ワークショップ(SWoPP 武雄 2005),情報処理学会研究報告 2005-HPC-103,pp.91-96,2005.8

122. 斎藤秀雄,田浦健次朗,近山隆 : 広域メッセージパッシングシステム用の遅延を考慮した接続管理. 2005 年並列/分散/協調処理に関する「武雄」サマー・ワークショップ(SWoPP 武雄 2005), 情報処理学会研究報告 2005-HPC-103,pp.181-185,2005.8

123. 堀田勇樹,田浦健次朗,近山隆 : 複数サブネット環境における自律的な故障検知機構. 2005 年並列/分散/協調処理に関する「武雄」サマー・ワークショップ(SWoPP 武雄 2005),情報処理学会研究報告 2005-OS-100,pp. 93—100,2005.8

124. 遠藤敏夫,田浦健次朗 : 高い耐遅延性を持つガウス消去法. 2005 年並列/分散/協調処理に関する「武雄」サマー・ワークショップ(SWoPP 武雄 2005), 情報処理学会研究報告 2005-HPC-103,pp. 121—126,2005.8

125. 白井達也,遠藤敏夫,田浦健次朗,近山隆 : 高いヒープ使用率のもとで高速なインクリメンタル GC. 2005 年並列/分散/協調処理に関する「武雄」サマー・ワークショップ(SWoPP 武雄 2005), 情報処理学会研究会資料 PRO-2005-2,2005.8 (持参原稿のみで予稿集なし)

126. 高橋慧,田浦健次朗,近山隆：マイグレーションを支援する分散集合オブジェクト. 2005 年並列/分散/協調処理に関する「武雄」サマー・ワークショップ(SWoPP 武雄 2005),情報処理学会研究会資料 PRO-2005-2,2005.8 (持参原稿のみで予稿集なし)

127. 田浦健次朗：Grid/P2P プログラミングモデルと関連技術. PPL Summer SchoolPPL Summer School 2005,東北大学,2005.9 (口頭講演のみで予稿集なし)

### (2) ポスター発表　　（国内　6 件、海外　10 件）

1. Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Sang-Zoo Lee and Jun'ichi Tsujii. GENIA corpus: A Semantically Annotated Corpus in Molecular Biology Domain. In the Proceedings of the ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001) poster session. pp. 68. 2001.
2. Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim and Jun'ichi Tsujii. The GENIA Corpus: an Annotated Corpus in Molecular Biology Domain. In the Proceedings of the 10th International Conference on Intelligent Systems for Molecular Biology (ISMB 2002) poster session. 2002.
3. Kumiko Tanaka-Ishii, Masato Yamamoto, and Hiroshi Nakagawa, Kiwi: A Multilingual Usage Consultation Tool based on Internet Searching, Proceedings of the Interactive Posters/Demonstrations, ACL-03, pp.105-108, Sapporo, 2003/7
4. Hong-Woo Chun, Tomoko Ohta, Jin-Dong Kim and Jun'ichi Tsujii. Building Patterns for Biomedical Event Extraction. In the The 15th International conference on Genome Informatics (GIW)-Poster and Software Demonstrations. 15(2). pp. 163. Universal Academy Press, INC. 2004.
5. Miura Kenji, Yoshimasa Tsuruoka and Jun'ichi Tsujii. Automatic acquisition of concept relations from web documents with sense clustering. In the IJCNLP 2004 Interactive Poster/Demo. pp.37-40. 2004.
6. Hidetaka Masuda, Shuuich Tsukamoto and Hiroshi Nakagawa, Recognition of HTML Table Structure, Proceedings of IJCNLP (International Joint Conference of Natural Language Processing) 2004, pp.183-188, Hainan-Island, China, March 2004/3
7. Hiroshi Nakagawa, Hiroyuki Kojima, and ra Maeda, Cinese Term Extraction from Web Pages Based on Compound word Productivity, 42nd Annual Meeting of the Association for Computational Linguistics (ACL2004), Third SIGHAN Workshop on Chinese Language Processing, pp.79-85, Barcelona, Spain, 2004/7
8. Minoru Yoshida and Hiroshi Nakagawa, Reformatting Web Documents via Header Trees, 43rd ACL2005 Poster/Demo Session, pp.121-124. Ann Arbor, 2005/7
9. Ayako Hoshino and Hiroshi Nakagawa, WebExperimenter for Multiple Choice Question Generation", HLT-EMNLP-05(Human Language Technology Conference Conference on Empirical Methods in Natural Language Processing, pp.18-19), Demo session, Vancouver, B.C., Canada 2005/10
10. Okanohara Daisuke. Partially Decodable Compression with Static PPM. In the Data Compression Conference 2005 poster session. 2005.
11. 堀田勇樹, 田浦健次朗, 近山隆: 分散環境における耐故障並列計算を支援する通信ライブラリ. 先進的計算基盤システムシンポジウム (SACSIS2004), May 2004.
12. 鴨志田良和, 田浦健次朗, 近山隆：複数クラスタ環境での並列 SAT-solver の実行. 先進的計算基盤システムシンポジウム (SACSIS2005),May 2005.
13. 斎藤秀雄, 田浦健次朗, 近山隆：動的スパニングツリーを用いた広域メッセージパッシングシステム用の集合通信. 先進的計算基盤システムシンポジウム (SACSIS2005),May 2005.
14. 白井達也, 遠藤敏夫, 田浦健次朗, 近山隆：高いヒープ使用率のもとで高速なインクリメンタル GC. 先進的計算基盤システムシンポジウム (SACSIS2005),May 2005.
15. 高橋慧, 田浦健次朗, 近山隆：マイグレーションを支援する分散集合オブジェクト. 先進的計算基盤システムシンポジウム (SACSIS2005),May 2005.

16. 遠藤敏夫, 田浦健次朗：高い耐遅延性を持つガウス消去法. 先進的計算基盤システムシンポジウム (SACSIS2005),May 2005.

## ５．３ 特許出願（国内　　0 件、海外　　件、その他　0 件）

なし

## ５．４ 受賞等

### (1) 受賞

1. 言語処理学会第 9 会大会 優秀発表賞、 2003.4、
山本真人、田中久美子、中川裕志、"検索エンジンに基づく多言語用例指南ツール:Kiwi"、発表番号 A1-5

2. Best Paper Award of PACLING'03、 Aug. 2003、
Kumiko Tanaka-Ishii、 Michiko Abe、 Hiroshi Nakagawa. "Categorization of movies using comments"、 PACLING'03 (Pacific Association for Computational LINGuistics)、 Halifax、 Nova Scotia、 Canada、 Aug. 2003

3. 山田雅信、 田浦健次朗、 近山隆、 高橋俊之：インクリメンタル PageRank による重要 Web ページの効率的な収集戦略　SACSIS 2004 学生優秀論文賞　2004 年 5 月

4. 大和 Adrian 賞: 英国 Royal Academy, 2004
マンチェスター大学と東京大学(辻井研究室)の「生命科学のためのテキストマイニング技術」の共同研究に対して、２００４年１１月

5. 堀田勇樹、田浦健次朗、近山隆：耐故障並列計算を支援する自律的な故障検知機構　SACSIS 2005　優秀若手研究賞　2005 年 5 月

6. Best Presentation Awards of WWW2005、
Kumiko Tanaka-Ishii (=presentor)、 Hiroshi Nakagawa (= co-author)、 "A Multilingual Usage Consultation Tool based on Internet Searching ? More than search engine、 Less than QA ?"、 Chiba、 Japan、 May 2005

7. Eclipse Innovation Grant(EIG)：米国ＩＢＭ
東京大学（辻井研究室）の「テキストマイニングのためのリソース構築と索引構造」の研究に対して、２００５年６月

### (2) 新聞報道

1. 田中久美子、中川裕志、Web 高度活用の時代へ――書いてない情報も Web から取り出せる、日本経済新聞社：NIKKEI NET「IT 羅針盤」、 2005/4/1

### (3) その他

研究代表者は、以下のような国際学会での基調講演、招待講演、チュートリアルに招待され、本プロジェクトの成果を報告した。

1. 招待講演, PSB01 Hawaii, 2001
2. チュートリアル講演, ISMB, Copenhagen, 2001
3. チュートリアル講演, HLT, San Diego, 2002
4. チュートリアル講演, Coling, Taipei, 2002
5. 招待講演、SCIE, Rome/Frascati, 2002
6. 招待講演, LangTech, Berlin, 2002
7. 基調講演, IEEE Int. Conf. of NLP and KE, Beijing, 2003
8. 招待講演, WS on Multi-layered Annotation, Bielefeld, Germany, 2004
9. 基調講演, LREC, Lisbon, 2004
10. 招待講演, EBI, Cambridge, 2004

11. 基調講演, ICON2004, 2004.
12. 基調講演, SMBM, Cambridge, 2005
13. 招待講演, BIT 2005, Tainan, 2005

### ５．５ その他特記事項

研究代表者の辻井は、２００５年７月より英国マンチェスター大学に付置された英国・国立テキストマイングセンターの Research Director に任命された。このセンターは、言語処理・テキスト処理の技術を広い分野の科学技術研究に適用することを目的に設立されたセンターであり、とくに、ＴＭ技術を生命科学へ適用することを当面の目標にしている。

このセンターは、ＧＲＩＤ技術を基盤とする e-Science 計画の一環を担っており、我々が本研究プロジェクトで目指した方向（スケーラブルなテキスト処理技術、オントロジー技術とテキスト処理技術の融合による有効な情報提供サービス、生命科学へのＴＭ技術の適用）ときわめて近い目標を持っており、本研究プロジェクトの研究代表者が、Research Director に任命されたことは、本研究の成果が広く世界に認められている証左と考えられる。

## ６　研究期間中の主な活動

### (1) ワークショップ・シンポジウム等

| 年月日 | 名称 | 場所 | 参加人数 | 概要 |
|---|---|---|---|---|
| 平成14年2月18日～20日 | Workshop on Natural Language Processing and Ontology Building for Biology | 東京ガーデンパレス | 59 名 | 生命科学の分野では、分野に関連する論文のアブストラクトが現在 1、000 万件以上テキストデータベースに収録されており、この行動に専門化された分野のテキストに、自然言語処理技術を適用して知識獲得を行うことが切望されている。<br>本ワークショップでは、言語と知識処理、特に知識獲得のためのテキストベースの研究の分野における世界の研究現況を把握するとともに、世界的に標準化の作業が進行する知識獲得のためのテキストデータベースの研究にも貢献していくために、分野の専門家との研究協力関係を結ぶことを目的とする。 |
| 平成14年2月20日 | 【公開シンポジウム】自然言語処理とゲノムオントロジー | 東京ガーデンパレス | 100 名 | 上記のワークショップに併催の公開シンポジウム。生命科学分野を対象とした自然言語処理やオントロジー構築に関して、国内外の代表的な研究者を演者として東京大学医科学研究所高木研究室などと共催で開催。 |

| | | | | |
|---|---|---|---|---|
| 平成16年<br>2月27日<br>～28日 | NTT コミュニケーション科学基礎研究所とのワークショップ | ラフォーレ修善寺 | 32 名 | 機械学習に関して先駆的な研究を行っている NTT と、言語の構造処理を中心に研究を行っている辻井 G とが研究成果を交換することで、相互の研究を促進することを目的とする |
| 平成16年<br>3月21日 | Beyond shallow analyses-Formalisms and statistical modeling for deep analyses | 中国<br>海南島 | 40 名 | 言語が示す規則性を、その構造的な側面と統計的な側面とにわけ、それぞれを数学的な枠組で捕えること、また、その２つの側面を統合したモデルを構築することは、今後の言語に関する数学的理論を考える上で重要な研究となっている。本ワークショップでは、欧米、アジアの代表的な研究グループおよび我々のグループから研究発表を行うことで、この分野の最新情報を交換すること目的としている。 |
| 平成16年<br>11月18日<br>～20日 | NTT コミュニケーション科学基礎研究所とのワークショップ | 浜名湖ロイヤルホテル | 24 名 | 機械学習に関して先駆的な研究を行っている NTT と、言語の構造処理を中心に研究を行っている辻井 G とが研究成果を交換することで、相互の研究を促進することを目的とする |
| 平成17年<br>3月11日 | e-バイオロジーイニシアティブ：新しい生物学の開拓 | 東京大学<br>武田ホール | 250 名 | バイオインフォマティクスの併催ワークショップとして、東大医科学研究所高木研究室などと共催で開催。 |

## (2) 招聘した研究者等

| 氏　名（所属、役職） | 招聘の目的 | 滞在先 | 滞在期間 |
|---|---|---|---|
| Mark Johnson<br>（米国ブラウン大学教授） | 記号的な言語処理手法と統計的な機械学習モデルとの統合に関する共同研究に関する打ち合わせを行う | 東京大学大学院<br>情報理工学系研究科<br>辻井研究室 | 平成16年<br>3月 6日～<br>4月 4日 |
| Aravind K. Joshi<br>（米国ペンシルバニア大学 教授） | TAG のための構文解析手法、および生命科学分野における文献情報処理に関する研究打ち合わせを行う | 東京大学大学院<br>情報理工学系研究科<br>辻井研究室 | 平成16年<br>3月15日～<br>3月26日 |
| Sofia Ananiadou<br>（英国立テキストマイニングセンター<br>副所長） | 「分子生物学の知識の構造化」に関する共同研究の発展のために、辻井 G で研究打合せを行う。また国際シンポジウム『e-バイオロジーイニシアティブ：新しい生物学の開拓』にて研究成果を発表してもらう | 東京大学大学院<br>情報理工学系研究科<br>辻井研究室 | 平成17年<br>3月 8日～<br>3月14日 |

| | | | |
|---|---|---|---|
| Udo Hahn<br>（ドイツ イエナ大学教授） | 「並列自然言語処理」に関する共同研究打合せのため。また国際シンポジウム『e-バイオロジーイニシアティブ：新しい生物学の開拓』にて研究成果を発表してもらう | 東京大学大学院<br>情報理工学系研究科<br>辻井研究室 | 平成17年<br>3月 8日～<br>3月12日 |

## ７　結び

**[研究目標からみた達成度]**

本研究の目標は、言語処理・知識処理・GRID ソフトウェア技術・エージェント技術などの成果を統合することで、膨大なテキスト情報の収集と処理、ユーザへの情報の提示に関する基盤技術を確立することであった。また、５年間の研究期間を理論・要素技術の研究を行う前半３年間と、成果の統合と総合化に重点を置く後半の２年間にわけ、メリハリのある研究を行うことを目指した。

この目標と戦略は、本プロジェクトが、理論・要素技術・統合システムのいずれにおいても世界をリードする成果をあげたことから、大きな成功であったと考えている。

プロジェクトメンバー間の協力も、ソフトウェアＧの GRID 技術、言語処理Ｇの文解析プログラム、オントロジーＧの生命科学分野に特化した辞書・オントロジーという成果を統合し、１４億語という巨大なテキスト処理を世界に先駆けて行う統合実験に結実するなど、きわめて円滑にであった。

**[研究成果と意義]**

本プロジェクトは、理論的な側面では、(1)非等号制約を認めるＭＥ、(2)素性構造文法の素性森を用いた確率モデル（３－１節)、(3)確率モデルを言語処理解析アルゴリズムへ組み込む一般的枠組み(３－４節)、という非常に幅の広い応用をもつ枠組みを開発した。これらの理論は、世界の有力な研究グループがシステム構築のために使用するなど、多くの論文で引用されている。

また、要素技術・研究資源の開発としては、(1)HPSG の高速・高耐性な文解析システム(Enju)の開発(３－１節)、(2)一般的な計算機環境で動作する高速クローラー（３－５節)、(3)分散計算機環境の有効使用を行う GRID ソフトウェア GPX（３－５節)、(4)高耐性な専門用語認識システム（言選）（３－８節)、(5)領域代数と素性構造のための索引構造とその検索系 MEDUSA（３－２節)、(6)生命科学のオントロジーと注釈つきのコーパスの作成[GENIA オントロジーとコーパス]（３－４節）という成果をあげた。これらは、いずれも、同種の要素技術で世界の最高水準のものである。

たとえば、Enju の性能は、処理速度では群を抜いて世界最高であり、精度の点でもエディンバラ大学・ケンブリッジ大学・PARC・ペンシルベニア大学という、この分野での先導的なグループのものと同等、あるいは、それ以上のものとなっている。また、GENIA コーパスは、規模と質の両面で生命科学の意味つきコーパスの最高のもとみなされ、世界で２４０超のグループによって使われている。

本プロジェクトの成功は、上記のような理論と要素技術に国際的な成果が得られたことだけではない。同様に重要なことは、これらの成果が有機的に連関されて、（３－７節）の統合実験、（３－８節）の統合サービスシステムなど、要素技術や理論に関する個別研究だけでは達成できない成果をあげたことである。

これらの成果は、プロジェクトの後半期、とくに、最終年度に得られたものであり、その外部からの評価はまだ定まっていない。今後、統合実験で得られた大規模な意味表現データや統合サービスシステムが、世界の研究者集団や実ユーザの生命科学者にどのように使われていくかが、プロジェクト評価としては重要となる。

プロジェクトリーダーとしては、米国流のアドホックな技術開発とは異なり、理論・要素技術から統合システムへという、Systematic で、かつ、Generic な基盤技術が確立できたことに満足している。また、深いテキスト処理を情報要求とは独立に事前に実行し、これをもとに情報要求に依存した知識処理・情報抽出を行うというプロジェクトの考え方が、実用上可能であることを実証できたことは、今後のテキスト・知識を中心にした情報システムの研究に大きな影響を与える。

**[今後の研究の展開]**

今回のプロジェクトでは、構造的に複雑な言語処理（単語列から意味構造を計算する文解析の処理、など）の技術が十分に実用的な性能を持って実行できること、また、その結果をオントロジーに基づく知識処理と結び付ける基本の技術を開発した。ただ、研究の主体は、単語列から意味構造を計算する過程に重点がおかれ、意味構造と知識処理とを結びつけて、より高度な推論処理を行うには至らなかった。この部分の研究を行うためには、

(1)テキストからの大規模オントロジーの（半）自動構築
(2)文境界を越えた文脈処理
(3)共通の大規模資源を共有することを可能にする分散計算機環境
(4)非記号的な推論形式と記号的推論形式を統合する技術

についてのさらなる研究が必要となる。今回の CREST 研究の成果は、これらのテーマのいくつかに焦点をてた発展研究に引き継がれる。

また、過去５年間の間に、Semantic Web、e-Science、Data Grid、生命科学における Text Mining という、本プロジェクトに関係が深い周辺分野が注目され、研究が加速している。本プロジェクトの成果は、テキスト処理の観点から、これらの分野に貢献することになる。

**[プロジェクトの運営]**

①　個別研究とプロジェクトの相互関係

プロジェクトの運営では、個人の研究者の独創性、理論的な緻密さが要求される理論・要素技術的な研究と、大規模な予算と集団研究でのみ可能となる統合的な研究とのバランスにもっとも注意を払った。

集団で行うシステム開発だけを目的にすると、浅い思いつきに基づいた、汎用性のないシステム開発に終わり、新規性・独創性・理論的整合性を求める研究者集団のプロジ

ェクトとしては不満足なものとなる。逆に、個人研究に重点を置きすぎると、相互の関連性が薄い、脈絡のない個別的な研究の集積となり、大規模な予算を使うプロジェクトとしての意義がなくなる。

この点で、理論・要素技術・大規模な資源構築に重点を当てる前期と、統合化を目指す後期というメリハリをつけた運営は、非常に良かったと思う。

② 研究費、若手研究者の育成

この種の研究では、優秀な人材の確保と、研究遂行のための資源構築が重要と考え、これらへの予算配分を最優先とした。韓国・ドイツからの若手研究者を含め、常に６－７名の常勤研究者を雇用した。また、GENIA コーパスの構築と分散計算機環境の整備というリソースの構築にも、人件費と同様に予算を使った。この結果、理論・要素技術・リソースのしっかりとそろった環境で基盤技術と統合システムの構築ができた。

また、プロジェクトが提供した人的・資源的な環境は、若手研究者にとっても刺激的な研究環境であり、彼らはそろって優れた成果を上げ、それぞれの分野で世界の研究をリードする研究者となっている。また、プロジェクト雇用の博士学生も、優秀な研究者が密度高くいるという環境で、多くの優れた成果をあげた。

③ プロジェクト組織の変更

プロジェクトの過程で、オントロジーG の中田助教授がドイツの教授、また、エージェント研究Gの西田教授が京都大学へとそれぞれ栄転され、それに伴う組織変更と研究目標の変更を行った。

オントロジーG のリーダーは、専門用語認識など、オントロジー研究の専門家である中川教授（東京大学）にお願いした。教授の興味は、テキストからのオントロジー構築であり、プロジェクトとの整合性もよく、交代は支障なく行われた。

また、前半終了時点での見直しで、西田教授の「非記号と記号の情報」、「身体性とインターフェース」の研究課題は、独創性のある、きわめて高い個別技術であり、京大栄転後も引き続きお願いすることとした。教授には、後期の統合的な研究では、プロジェクト打ち合わせ会で積極的な助言をいただくにとどめ、個別要素研究を引き続きお願いすることとした。この判断は、エージェントG の個別研究成果の質の高さと、その研究が今後のテキスト処理での HCI に与える影響の大きさを考えると、正しいものであった。

**[研究成果の普及と宣伝]**

本プロジェクトでは、外の研究者集団とも緊密な連携をもち、成果を外部に広めることに留意した。とくに、生命科学者とは、共同研究や国際ワークショップの共同開催（２回）など、緊密に連携をとった。結果として、生命科学のためのテキストマイニングという新しい分野に世界の研究者の関心を集めることに成功した。また、２００４年度に開催した「深い文解析」に関するワークショップも、本プロジェクトの主張の一つである「きっちりとした言語学理論に基づく言語処理」に焦点をあて、これまで独自に研究を行っていたペンシルベニア大学・Ｐａｒｃ・エディンバラ大学などの研究グループを共通の研究目標を持った集団にすることに貢献した。