戦略的創造研究推進事業
発展研究（SORST）

研究終了報告書

研究課題
「次世代テキストマイニングの
技術基盤に関する研究」

研究期間：平成１７年１１月　１日～
平成１９年　３月３１日

辻井　潤一
（東京大学、教授）

１．研究課題名
次世代テキストマイニングの技術基盤に関する研究

２．研究実施の概要

２．１ 基本構想
[研究の目標]
これまでのテキストマイニング(TM)の技術は、テキストを単なる単語集合とみなして、これに確率・統計モデルに基づくマイニング適用する量の技術であった。これに対して、本研究では、文の統語・意味構造、テキストの文脈構造、および、陽には表現されない背景知識を取り扱う質の技術に焦点をあて、これを量に基づく技術に統合することで、従来の技術をはるかに凌駕するＴＭの技術基盤を確立する。また、膨大なテキストの集積と複数分野での知識統合が進展し、TM 技術への需要が顕在化している生命科学での特定タスクを取り上げ、開発したＴＭ技術の有効性を実証する。
[研究の背景と着眼点]
従来のＴＭ技術が、テキストを単語集合(Bag of Words: BOW)とみなす技術にとどまっていたのは、（１）大量の言語データを処理するのに十分な処理効率と耐性とを備えた言語処理の技術、（２）意味処理・知識処理に必要となる大量のリソース（意味辞書、知識ベース）の構築、（３）質・量ともにスケール・アップした処理を支える計算機インフラの開発、の３つの技術分野が未成熟であったためである。
しかしながら、本発展研究に先行する CREST 研究で示したように、構造・意味を取り扱う処理手法と量的側面を扱う機械学習からの処理手法との融合により、大量・広範なテキストを処理するのに必要な高耐性・高効率な言語処理技術は、現時点で十分達成可能なものとなりつつある。また、大量テキストからの意味辞書・知識ベースを構築する半自動的な手法の発展により、生命科学をはじめとするいくつかの専門分野では、現実に膨大な言語・知識リソースの構築が開始されている。さらに、処理と記憶の能力を格段に向上させる GRID 技術の発展も著しい。本研究は、これら３つの分野での成果を統合することで、BOW モデルでの限界を突破する、全く新しいＴＭ技術の基盤を確立する。

２．２　研究の実施経過
本プロジェクトは、上記の基本構想のもとで３年間の計画で平成１７年１１月より開始された。しかしながら、平成１８年９月から開始された５年間のプロジェクト、文部科学省研究補助金・特別推進研究「高度言語理解のための意味・知識処理の基盤技術に関する研究」との重複が大きいとの判断から、本発展研究の辞退を申し入れ、平成１９年度からの辞退が受理された。３年計画の研究が１年５ヶ月に短縮し、また、平成１８年９月からは、上記の特別推進研究が開始することから、研究総括・三谷教授（北陸先端大）との綿密な協議の上、研究計画を以下のように変更した。

（１）テキストからの知識構築、分野知識に基づく言語処理など、基礎的な色彩の強い研究項目、および、大規模テキストマイニングのための並列分散の計算機環境の構築など、インフラ構築の研究項目を発展研究から切り離し、特別推進研究へと移行する。
（２）発展研究では、テキストマインニングの生命科学への適用に焦点を当て、生命科学者が研究現場で使用できるシステムの構築、また、そのために必要となる基盤の言語処理技術の確立に焦点を当てる。
（３）上記の変更に伴い、意味・知識の処理を担当していたオントロジグループ（中川浩志教授・東京大学)、計算機インフラの構築担当の広域ソフトウェアグループ（米澤明憲教授）が本研究の主体からは抜けることとし、本研究は研究代表者が直接統括する言語処理グループが上記（２）の研究項目を推進する。これに伴い、平成１８年度予算の相当額を減額することとなった。

このような大幅な研究期間の短縮と研究項目の変更にも関わらず、研究の実施はきわめてスムーズに行われた。特に、生命科学者のための統合システムとそのための基盤的な言語処理技術の開発という、短期間での目標設定を行ったことにより、焦点を絞った技術開発を行うことができた。具体的には、次の３つの分野で非常に優れた成果を挙げた。

（1） 生命科学の課題解決型システムの開発：
先行する CREST の終了時点でそのプロトタイプが完成した２つの統合システム(Info-Pubmed、MEDIE)を、実ユーザ（国立遺伝学研究所、理化学研究所、マンチェスター大学バイオ研究センターなどの生命科学者）からのフィードバックに基づいて個別課題のための機能を拡充した。　特に、病疾患と遺伝子との関係をテキストからマイニングするシステム、および、１５００万件の Medline テキストベースから、蛋白質相互作用に関する情報抽出を行うシステムを構築した。
高速・高精度な分野適応型言語処理技術の開発：
本研究の目標は、従来の BOW モデルの限界を突破する言語処理に基づくテキストマイニング技術の開発であった。このためには、大量のテキスト集合を高速・高効率、かつ、高精度で処理できる言語処理技術の開発が不可欠である。我々が本研究で開発した英語解析システム(Enju)は、本格的な言語理論(HPSG)に基づき、１５００万の論文抄録（７０００万文、１４億語）という巨大なテキスト集合を処理し、かつ、１文あたりの処理時間が、１５msec という高速なものである。また、MEDLINE に特化した統計モデルを学習する領域適応の技術を使うことで、F-値が９０％を超える精度が達成できることを実証した。（1）の課題解決型システムは、すべて、この英語解析システムを使った成果である。
（3）知識・言語リソースの構築：
先行する CREST プロジェクトで開発したアノテーション・コーパス（GENIA）は、生命科学におけるテキストマイニング技術を開発するための言語資源として世界の研究者に使われている。本研究では、この GENIA コーパスを知識・意味処理のための基礎コーパスとして整備するために、これまでの名詞概念のアノテーションを動詞的概念（事象）に拡張し、１０００論文抄録に対する生命事象アノテーションを完成した。また、GENIA オントロジを生命科学の主要なオントロジ（GO、Mesh、BioPax）とリンクし、リソース間の相互利用を可能とした。

２．３ 研究成果
実施経過の項にあげた３つの研究分野の成果を以下に整理する。

（1） 生命科学の課題解決型システムの開発

（1－1）意味に基づく検索システム（４．１．１節）：
英語解析システム(Enju、４．２．１、４．２．２、４．２．３参照)、先行する CREST で開発したＮＥＲ(Named Entity Recognizer)、領域代数に基づくテキスト索引システム（４．２．５参照）を統合することで、言語処理結果を活用した知的検索システムを構築した。このシステムは、国立遺伝学研究所、マンチェスター大学バイオ研究センターの研究者によって実際に使用されている。
（1－2）病疾患と遺伝子の関係マイニングシステム（４．１．２節）：
病疾患、遺伝子の辞書を整備し、CREST で開発したＮＥＲの精度を向上させること、また、英語解析システム(Enju、４．２．１、４．２．２、４．２．３参照)の結果を素性として使うＭＥ(Maximum Entropy)による分類器を使うことで、７０％を超える精度のマイニング結果を得る。このシステムは、産業総研の生命科学グループが開発する LEGEND システムの一部として組み込まれて、実際の研究場面で使われている。
（1－3）蛋白質相互作用の情報抽出（４．１．３節）：
英語解析システム(Enju、４．２．１、４．２．２、４．２．３参照)、ＮＥＲの成果を統合することで、PDF 形式の論文から、蛋白質相互作用が記述されている箇所を同定するシ

ステムを開発し、ヨーロッパ・アメリカが推進している国際的なコンペティション(BioCreative)に参加し、１５の参加チームの中で３位以内の成績をあげる。

（２） 高速・高精度な分野適応型言語処理技術の開発

（２－１）構文解析器の高速化（４．２．１節）:
CREST 終了時では、我々の英語解析システム(Enju)は、MEDLINE 一文の処理に 平 均１秒の処理時間であった。この速度は、言語理論に基づく解析器としては、世界でトップであったが、CFG フィルタリング、Super Tagging、Shift-Reduce 型決定的解析手法を用いることで、精度の劣化なく１００倍以上の高速処理が可能なことを実証した。
（２－２）混合型解析モデルによる高精度化（４．２．２節）:
Shift-reduce 型の解析器と統計的分類器をの活用で、比較的高い精度の依存構造解析器が構築できる。このような依存構造解析器を HPSG 解析の前処理に使うことで、８６．３%の解析精度を８８．２%にまで向上させることを示した。これは、CFG の浅い解析系の精度とコンパラブルな精度であり、HPSG がより深い構造や CFG では対処不可能な構造を解析できることから、我々のアプローチの実用的な優位性が示せた。
（２－３）解析器の分野適応（４．２．３節）:
特定の分野のコーパスで学習された処理系は、他の分野に適用されると、(時としては、大きく）その精度が劣化する。この研究では、我々の解析器(Enju)がこの点でも優れた特性を持ち、これを活用することで小規模なコーパスでの適応学習が可能であることを示し、４万文を超えるコーパスを使ったモデルを８０００文程度で適応させ、劣化を防ぎ、８６．３%の精度を９０．１%に向上させることに成功した。この結果は、我々の成果を広い分野で活用する基礎技術であり、工学的な価値が大きな成果である。
（２－４）POS Tagger の分野適応（４．２．４節）:
品詞タギングは、英語解析、ＮＥＲ、ＥＲなどの処理の前処理として使われ、ここでの誤りが後続処理の誤りに拡大されることから、精度の向上が必須のものとなっている。我々は、タグセットが大きくなっても速度劣化のない CRF(Conditional Random Field)を開発し、現時点では、世界でもっとも高い精度の Tagger を開発、また、能動学習により非常に少数のコーパスで分野適応可能なことを示した。
（２－５）領域代数を使ったテキスト検索システム（４．２．５節）:
巨大テキスト集合の検索の主流は、単語や文字の並びを対象とする Full Text Search である。これは、言語処理の結果を蓄積し、索引構造に反映する有効な技術が存在しなかったことが理由である。この研究では、ウォータルー大学の領域代数を、埋め込み構造と変数つき構造に拡張し、１５００万抄録の巨大なテキスト集合に対する Enju や NER の処理結果を蓄積し、検索するシステムを構成した。現在の XML データベースでは、１００分の一程度のデータも蓄積できず、検索速度もきわめて遅いことを確認している。このシステムは、MEDIE（４．１．１節）の基盤システムとして使われている。

（３）知識・言語リソースの構築

（３－１） GENIA コーパスの構築（４．３．１節）:
上記（１)、(２）の技術開発を系統的に行うためには、その基礎資料となるアノテーションつきのコーパスの構築が不可欠である。我々は、生命科学のテキストマイニングのために、先行する CREST プロジェクトから GENIA コーパスの構築を行ってきた。本研究では、共参照関係を新たに加え、統語構造とＮＥのアノテーションを見直すことで、さらにその質を向上させた。コロラド大学の調査では、GENIA コーパスは、他のコーパスに比べて圧倒的に使用グループが多いコーパスとなっている。
（３－２）GENIA Event アノテーションの設計と構築（４，３，２節）:
GENIA のこれまでのアノテーションは、米国の MUC などで行われたアノテーション作業を生命科学に適用したもの(POS、統合構造、共参照)か、生命科学固有であっも構造的は単純なもの(NE)であった。これをさらに一歩進めて、生命科学固有で、かつ、構造的にも複

雑なアノテーションに事象のアノテーションがある。本研究では、１０００抄録の事象アノテーションを完了したが、このような大量で、かつ、複雑な事象アノテーションは世界的に最初のものとなっている。

[**注意**]
研究の実施経過（前項[B]を参照）でも述べてように、本発展研究は、開始後8ヶ月を経過した時点で特別推進研究とのかかわりで、研究組織を変更し、オントロジグループおよび広域ソフトウェアグループが移行することになった。したがって、本項では、1年5ヶ月間の研究を中核的に行った言語処理グループのものに限って記述した。

３．研究構想

２．２で述べたように、本研究は３年計画で発足したが、科学研究補助金・特別推進研究が平成１８年９月から開始することになったために、平成１９年３月で終了する１年５ヶ月のプロジェクトとなった。また、平成１８年度９月からは、基礎研究、インフラ構築に関与する研究を特別推進に移行し、当初の３グループによるプロジェクト構成を改め、１グループによる研究とした。

これに応じて、研究計画、予算などが変更されたが、本報告では、まず、３．１節でプロジェクト発足時の計画を記述し、３．２節で当初計画からの変更点をまとめる。

３．１ 研究発足時の研究構想

３．１．１　研究のねらい

これまでのテキストマイニング（TM）の技術は、テキストを単なる単語集合とみなして、これにマイニング適用する量の技術であった。これに対して、本研究では、文の統語・意味構造、テキストの文脈構造、および、陽には表現されない背景知識を取り扱う質の技術に焦点をあて、これを量の技術に統合することで、従来の技術をはるかに凌駕するＴＭの技術基盤を確立する。また、膨大なテキストの集積と複数分野での知識統合が進展し、TM技術への需要が顕在化している生命科学での特定タスクを取り上げ、開発したＴＭ技術の有効性を実証する。

３．１．２　研究の内容

本研究は、（A）文の構造から意味への基盤技術、（B）意味から知識への基盤技術、（C）大量で複雑な計算を支える計算インフラの構築、という３つの項目に分けて進める。ただ、これらの研究項目は、生命科学でのＴＭという実応用での統合を意識し、緊密な連携のもとに実施する。最終年度には、統合システムの構築と運用を行う。

（A）文解析技術（構造から意味への技術）：
文解析システム(Enju)の意味処理機能の強化、および、汎用 TM 技術のための分野適応性の研究を行う。すなわち、

(1-1)意味が変容した言語構造（事象名詞、かん喩表現、など）の意味表現への変換処理

(1-2)分野依存性をコーパスから学習する適応的な文解析システム

の研究により、現在技術の限界（係り受け関係単位で 90%以下）を大幅に改善（95%以上）するための研究を行う。とくに、分野に特化した意味体系を構造、意味処理に活用するモデル、少量の構造付きコーパスからの確率モデル構築を行う方式を確立する。

（B）文理解技術：
（１）の意味構造を機械学習の入力とし、複合的な知識単位（事象、過程）を認識する手法の研究を行う。

(2-1) 複雑な語構成を持った固有表現の認識（NER）の機械学習による適応的な手法

(2-2) 複合単位（句、節）の意味表現から分野依存な事象(Event)・過程(Process)を認識する手法

(2-3) 文境界を越えた共参照関係の認識処理

(2-4) 我々の持つ生命科学のコーパス GENIA(１００万語)に、事象・過程に関する生命科学者のアノテーションを加え、言語と知識の関係を系統的に研究するための基礎コーパスを構築する。

（C）計算リソース（Grid 環境の開発）：

本格的なTM技術のために､分散的なリソース管理と活用､強大な計算能力を提供するGrid環境の開発を行う。

(3-1) 知的検索のための意味にもとづくテキスト索引構造の開発

(3-2) 巨大テキスト集合の構造解析、TM処理を実行するための計算機リソースの構築

（D）統合システムの開発：

生命科学でのTMの典型として、次の２つの課題を取り上げ、統合システムとして構築する。

(4-1)蛋白質相互作用の自動抽出：文の意味構造、事象構造を使うことにより、単語列ウィンドウでの抽出結果よりもはるかに高い精度が得られることを大規模実験で実証する

(4-2)病疾患－遺伝子ネットワークの自動抽出：事象構造を素性とする機械学習により、病疾患と遺伝子の関係、とくに、研究者集団にとっても未知な関係を発見するシステムを開発する。

３．１．３　研究の主なスケジュール

項目	平成17年度 (5ヶ月)	平成18年度	平成19年度	平成20年度 (7ヶ月)
事象アノテーションの仕様**	←	→		
事象アノテーション*		←	→	
共参照アノテーションの仕様**	←	→		
共参照アノテーション*		←	→	
GENIAオントロジの検討**	←	→		
GENIAオントロジ・辞書		←	→	
意味解析（分野適応）*		←	→	
意味構造の設計**	←	→		
ＮＥＲ*	←	→	→	
事象認識*		←	→	
並列プログラミング言語の開発	←	→		
負荷分散アルゴリズムの研究	←	→	→	
統合システム１** （プロトタイプ）	←	→		
統合システム１ （本格システム）			←	→

統合システム２** （プロトタイプ）		←――→		
統合システム３ （本格システム）			←――→	←――→
ターミノロジー抽出	←――→	←――→		
個別領域コーパスからの知識抽出		←――→	←――→	←――→
半構造テキストからのマイニング	←――→	←――→		

３．１．４　研究の実施体制（研究チームの構成）

言語処理グループ
研究実施項目：言語の意味・知識処理に関する技術開発と言語・知識リソースの構築
概要：意味・知識の処理を導入することにより、高い精度の構文解析技術を開発し、言語の構造解析のための確率モデルとその分野適応性の研究を行う。また、言語の意味構造を知識へと写像するための技術の開発、とくに、そのための機械学習の研究を行うとともに、研究に必要なリソース（意味・知識をアノテートしたコーパスの作成、分野オントロジの整備）を行う。
研究の中核グループとして、ほかの２グループの研究の進行管理、統合実験の準備と実施。生命科学の研究グループや海外の協力グループとの連携をとる。

広域ソフトウェアグループ
研究実施項目：テキストマイニングの並列処理を支援するシステムに関する研究
概要：大規模なデータに対するテキストマイニングは、大量の計算能力(CPU)とメモリ、ストレージの全てを必要とする。広域ソフトウェアグループではそれを安価なクラスタ型計算機で、耐故障性を持って、効率よく実行できる並列プログラミング言語(既存スクリプト言語の拡張)と、より簡易に利用できるエンドユーザツールを研究・開発する。その枠組みの元で、CPU およびネットワーク帯域を資源として考慮した負荷分散(資源割り当て)アルゴリズムについて研究する。

オントロジグループ
研究実施項目：言語知識マイニングシステムに関する研究
概要：個別学問ないし技術領域の用語集合、すなわちオントロジを具体的なアプリケーションで利用できる言語知識として構築する方法およびそのためのマイニングツールに関する研究を以下の観点から実施する。
（１）専門領域からのターミノロジー抽出：
すでに開発した用語抽出システム「言選 Web」は複合語の構造を利用するため、例えば単独の Web ページのような小コーパスでは性能が良いことが既に立証されている。ここでは、コーパス規模に対する用語抽出の性能の変化を分析し、大規模コーパスにも効果を発揮する柔軟な方法を検討する。また、CREST においても進めていた用語抽出システムの多言語化についても進展させる。
（２）個別領域コーパスからの知識抽出：
既存の用例検索システム Kiwi を個別領域のコーパスに対応させる。コーパス対応型の Kiwi を利用して、個別分野からの高速用例抽出システムを実現し、それを足がかりにして実用的な言語知識を抽出する。具体的には、a)対象領域における因果関係を表す言語知識のマイニング、b)日英特許対訳コーパスから、個別技術分野の特許翻訳に特有の言い回し、定

型表現の対訳知識の獲得、に関する研究を行う。
（３）半構造テキストからのマイニング：
html やＸＭＬのようないわゆる半構造データからの言語知識獲得は、地のテキストに対する自然言語処理とレイアウト分析から得た構造知識を総合して行うと効果的である。このための手法、システムの開発、Web データへの応用について研究する。

３．2 計画の主要な変更点

特別推進研究の発足が内定した時点（平成１８年６月）で、研究総括・三谷教授（北陸先端大）と協議し、広域ソフトウェアグループ、オントロジグループの研究は、基礎研究、インフラ構築の研究であり、長期にわたる研究であるので、特別推進に移行する。したがって、以降の研究は、言語グループが遂行し、特に、発展研究としてのまとまりをつけるために、平成１９年３月までに生命科学者に提供する統合システム（３．１．２節（D））を完成させることに努力を傾注する。　また、このために必要な研究項目である、生命科学分野のコーパス構築とそれを利用する分野適応型の言語処理技術、の開発を行う。
すなわち、３．１．２節の研究項目のうち、
（A）文解析技術（構造から意味への技術）：
(1-2)分野依存性をコーパスから学習する適応的な文解析システム
（B）文理解技術　：
(2-4) 我々の持つ生命科学のコーパス GENIA(１００万語)に、事象・過程に関する生命科学者のアノテーションを加え、言語と知識の関係を系統的に研究するための基礎コーパスを構築する。
（D）統合システムの開発：生命科学での TM の典型として、次の２つの課題を取り上げ、統合システムとして構築する。
(4-1)蛋白質相互作用の自動抽出：文の意味構造、事象構造を使うことにより、単語列ウィンドウでの抽出結果よりもはるかに高い精度が得られることを大規模実験で実証する
(4-2)病疾患－遺伝子ネットワークの自動抽出：事象構造を素性とする機械学習により、病疾患と遺伝子の関係、とくに、研究者集団にとっても未知な関係を発見するシステムを開発する。
を発展研究で実施し、それ以外の項目を特別推進研究に移行することとした。

３．１．３節の研究スケジュールのうち、**と*の項目が本研究で実施した項目、**は、本研究の期間内で終了したもの、*は、その成果を特別研究が引き継ぐ項目を示している。

３．３　研究進行中の問題店と新たに生まれた研究方向

大きな計画変更が行われたが、計画は、次節の研究成果が示すように、きわめて順調に実行できた。これは、成果の切り分けが発展研究と特別推進の間で行われただけで、実態上の研究体制はそのまま維持できたことによる。
発展研究の終了時期に合わせて統合システムが完成したことは、今後、機能向上をおこなったサービスを組み込むプラットフォームが整備できたことを意味し、将来の研究にとっても有意義であった。
また、新たな機能の追加を許すテキストマイニング用のプラットフォームという考え方は、IBM のＵＩＭＡコンソーシアム、e-Science におけるワークフロー・ソフトウエアの利用など、ここ２年の間に大きく広がってきた。
我々の研究グループでも、CREST/SORST で開発してきたツール群を UIMA　Ｃｏｍｐｌｉａｎｔな形式に移行しており、これを基礎に上述の統合システムを再構築することで、ほかの研究グループが開発したツールを自由に Plug-In できる Inter-operable な Software 環境の典型を作り出すことができると考えている。日本から発信する、Ｉｎｔｅｒｏｐｅｒａｂｌｅなソフト環境で世界をリードしていきたいと考えている。

４．研究実施内容

４．１　“BioNLP とテキストマイニングの統合型システム／アプリケーション”

４．１．１　MEDIE ― 意味構造を利用した関係概念検索システム

はじめに

近年、膨大な知識がテキストという形で蓄積されるようになり、テキストから効率的に情報を獲得する手法が注目されている。同時に、文書ではなくより詳細な情報単位の検索に対する要求も増加している。例えば、生物医学分野の網羅的な文献データベースである MEDLINE には約 1,500 万件の文献が登録されており、その数は年 10% のペースで増加し続けている。この膨大な文献から、遺伝子-疾患相関やたんぱく質相互作用などの**関係概念**を検索することが必要とされている。しかし、関連する文献を全て読むのは困難であり、一方、単純な文書検索ではこのような詳細な情報を高精度で特定することができない。

本節では、MEDLINE から関係概念を高精度かつ高速に検索するシステム **MEDIE** (Miyao et al. 2006)について報告する（図 1)。本システムは、テキストの意味構造を計算する**オフライン処理**と、ユーザの入力を受け付け、意味構造を検索する**オンライン処理**から構成される（図２)。

- オフライン処理：HPSG 構文解析器(Miyao et al. 2005)と用語認識器(Tsuruoka et al. 2004)を適用し、述語項構造とオントロジ ID が付与された**意味構造付きテキスト**を作成する。
- オンライン処理：ユーザの検索要求を拡張領域代数(Masuda et al. 2003)のクエリに変換し、拡張領域代数の検索アルゴリズムを用いてクエリの構造とマッチする意味構造を検索する。

本システムでは、意味構造を指定することにより詳細な情報単位を検索することができ、また主な処理はオフラインで行うため高速な検索が可能である。以下では、MEDIE システムの概要と評価実験について詳述する。

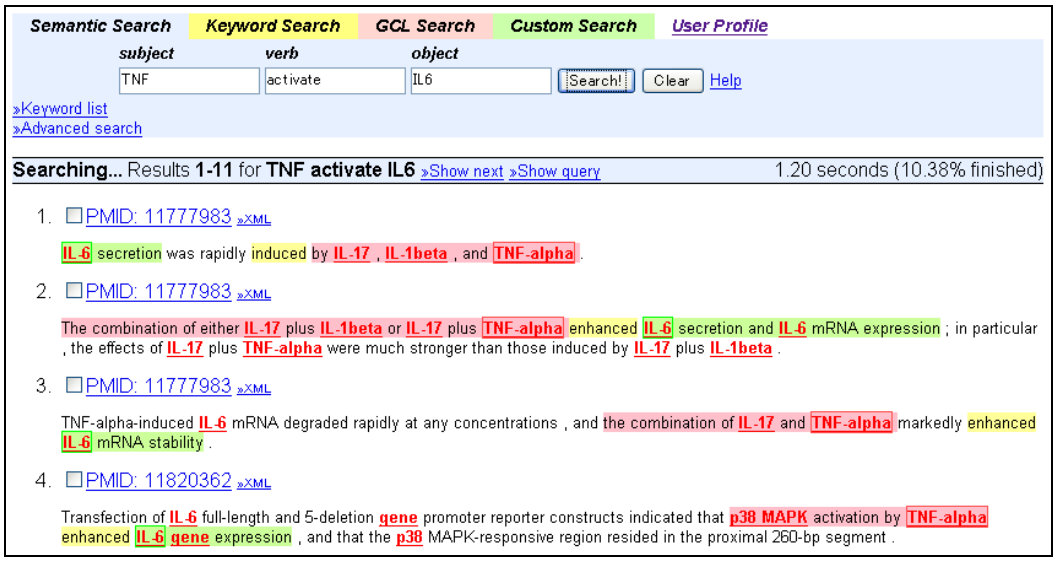


図 1：MEDIE

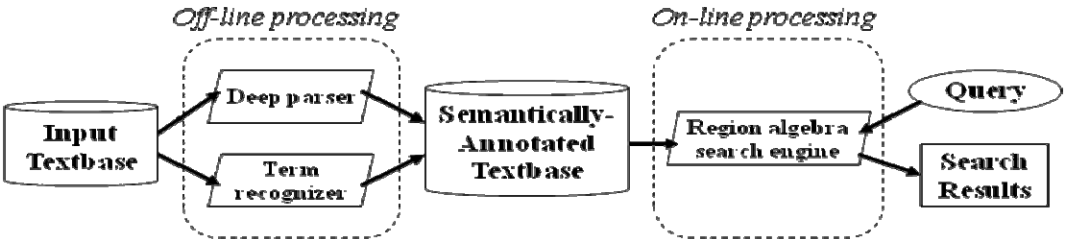


図 2：MEDIE のアーキテクチャ

HPSG 構文解析

HPSG は、文とその意味との対応関係を系統的に記述する現代文法理論の一つである。従って、HPSG に基づく構文解析により、句構造だけでなく述語項構造などの意味構造を計算することができる。述語項構造は文中の単語間の意味的関係を表現する意味構造である。本研究では、我々が開発した高精度かつ高速な HPSG 構文解析器(Miyao et al. 2005)を MEDLINE に適用した。さらに、MEDLINE 全文を現実的な時間で解析するため、東京大学の 2 キャンパスにまたがる 170 ノードのクラスタ(340 Xeon CPU)を GXP (Taura 2004)を用いて利用した。これにより、MEDLINE 全体の HPSG 構文解析を 9 日間で完了した(Ninomiya et al. 2006)。

表 1：遺伝子・タンパク質オントロジ

Symbol	CRP
Name	C-reactive protein, pentraxin-related
Species	Homo sapiens
Synonym	MGC88244, PTX1

専門用語認識

近年、特に生物医学分野では、大規模なオントロジの構築が行われている。オントロジはある分野の用語を系統的に整理したデータベースであり、テキスト表現をそれが指す実世界の実体へ写像するものと考えることができる。例えば、遺伝子・たんぱく質オントロジである GENA のエントリを表 1 に示す。 “Symbol”, “Name”, “Species” はそれぞれ短縮名、学名、生物種を表す。“Synonym”は同じたんぱく質または遺伝子を指す別名を表す。この表から、CRP, MGC88244, PTX1 などが同一の実体を指していることが分かる。テキストに現れる用語に対してオントロジ ID を付与することで、テキスト表現の差異を標準化し、用語の示す意味を表現することができる。本研究では、辞書ベースの用語認識アルゴリズム(Tsuruoka et al. 2004)を適用し、テキスト中の専門用語（遺伝子・たんぱく質、病名）に対してオントロジ ID を付与した。

また、関係概念を記述するには動詞的表現も重要である。例えば、たんぱく質の活性化を表す表現としては“activate”や“enhance”などが考えられる。このようなイベントを表す表現の種類は専門用語よりはるかに少ないが、オントロジが存在しないため、本研究においてイベント表現オントロジを作成した。まず MEDLINE の 500 アブストラクトを分析し、頻出する 167 のイベント表現を 18 のイベント型に分類した。これを用いて、専門用語の認識と同様にテキスト中のイベント表現に対してオントロジ ID の付与を行った。

拡張領域代数

本研究では、タグのネスト及び交差を含む構造化テキストの検索のために拡張領域代数(Masuda et al. 2003)を適用した。拡張領域代数は、テキスト上の領域（単語列）を引数とする演算子で定義される（表 2）。A と B は領域を表し、演算の結果も領域である。4 つの包含演算子（>, >>, <, <<）は XML における先祖／子孫関係を表す。例えば、“A > B”は B を子孫にもつ領域 A を表す、これらの演算子で記述するクエリによる検索は、最初の解の検索は定数時間、全解検索はクエリ中の単語の最低頻度に比例する時間で抑えられる。

表 2：拡張領域代数

[tag]	“<tag>”タグで囲まれた領域
>	B を含む領域 A
>>	B を含む領域 A(A はネストしない)
<	B に含まれる領域 A
<<	B に含まれる領域 A(B はネストしない)
&	A と B を含む領域
|	A または B を含む領域

例えば、“*CRP excludes something*”の意味構造を検索するクエリは、

[sentence] >>(([word arg1="$subject"] > exclude) & ([phrase id="$subject"] > CRP))

と表すことができる。このクエリは、文が単語“*exclude*”を含み、かつその第一引数(“arg1”)の句が単語“*CRP*”を含むことを表している。述語-項関係は変数“$subject”によって表されている．拡張領域代数クエリをユーザが直接入力するのは難しいため、本システムでは、ユーザが主語 x, 目的語 y, 動詞 v を入力すると、x, y, v をオントロジ ID に置き換えた上で、拡張領域代数クエリを自動生成する。

表 3：意味構造付き MEDLINE のサイズ

文献数	14,785,094
アブストラクト数	7,291,857
文数	70,935,630
単語数	1,462,626,934
構文解析に成功した文の数	69,243,788
述語-項関係の数	3,094,105,383
用語数（遺伝子）	84,998,621
用語数（たんぱく質）	27,471,488
用語数（病名）	19,150,984
用語数（イベント表現）	51,810,047
MEDLINE のサイズ	9.3 GByte
意味構造付きテキストのサイズ	292 GByte
インデックスのサイズ	954 GByte

表 4：実験に用いた検索要求

1	*something* inhibit ERK2
2	*something* trigger diabetes
3	adiponectin increase *something*
4	TNF activate IL6
5	dystrophin cause *disease*
6	macrophage induce *something*
7	*something* suppress MAP phosphorylation
8	*something* enhance p53 (negative)

評価実験

評価実験に用いたデータは２００４年度末時点で MEDLINE に登録された全文献約 1,500 万件である。これらにまずオフライン処理を行い、意味構造付き MEDLINE を作成した。表３に処理前・後のデータサイズを示す。また、実験に使用した検索要求を表４に示す。“*something*”は任意の語を、“*disease*”は病名、“(negative)” は否定文を検索することを意味する。これらは生物学者により選択された生物医学分野における典型的な検索要求である。

検索件数、検索時間、検索精度を表５に示す。実験では、述語項構造を利用するか、オントロジ ID を利用するか、の 2 点を変化させ、それぞれが検索結果に与える影響を調査し

た。“キーワード検索”は入力単語の共起のみの検索、“意味構造検索”は述語-項関係を利用した検索であり、この比較により述語項構造を利用する効果を検証する。クエリ番号の接尾辞はオントロジ ID を利用するかどうかを表す。X-1 はオントロジを利用せず、X-2 は遺伝子、たんぱく質、病名オントロジのみ用い、X-3 は全てのオントロジを利用する。検索精度は適合率とキーワード検索の結果に対する相対再現率を測定した。精度測定には、各クエリの検索結果 100 件を上限として、全ての結果をランダムに結合したもの(1,839 文)を生物学者が正解判定したものを用いた。

実験結果から、意味構造検索により適合率が有意に改善したのが分かる。多くの検索で適合率は 80%を越え、検索要求 4 と 5 についてはほぼ 100%である。これは、述語項構造を利用することで関係概念が正確に特定できていることを示している。また、検索件数の増加からオントロジ利用の有効性も示された。特に、イベント表現オントロジが有効であることが分かる。

表 5：検索件数、検索時間、検索精度

クエリ番号	キーワード検索			意味構造検索			
	検索件数	検索時間 (first/all)	適合率	検索件数	検索時間 (first/all)	適合率	相対再現率
1-1	252	0.00/1.5	74%	143	0.01/2.5	96%	69%
1-2	348	0.00/1.9	61%	174	0.01/3.1	89%	69%
1-3	884	0.00/3.2	50%	292	0.01/5.3	91%	42%
2-1	125	0.00/1.8	45%	27	0.02	85%	38%
2-2	113	0.00/2.9	40%	26	0.06/4.0	85%	48%
2-3	6529	0.00/12.1	42%	662	0.01/1527	76%	19%
3-1	287	0.00/1.5	20%	30	0.05/2.4	80%	30%
3-2	309	0.01/2.1	21%	32	0.10/3.5	81%	29%
3-3	338	0.01/2.2	24%	39	0.05/3.6	82%	33%
4-1	4	0.26/1.5	0%	0	2.44/2.4	－	－
4-2	195	0.01/2.5	9%	6	0.09/4.1	83%	22%
4-3	2063	0.00/7.5	5%	94	0.02/10.5	95%	40%
5-2	287	0.08/6.3	73%	116	0.05/14.7	97%	51%
5-3	602	0.01/15.9	50%	122	0.05/14.2	96%	46%
6-1	10698	0.00/42.8	14%	1559	0.01/3015	65%	71%
6-3	42106	0.00/3380	11%	2776	0.01/5100	61%	45%
7	87	0.04/2.7	39%	15	0.05/4.2	67%	29%
8	1812	0.01/7.6	19%	84	0.20/29.2	87%	37%

まとめと今後の課題

本節では、MEDLINE のための高精度かつ高速な関係概念検索システム MEDIE について報告した。まず HPSG 構文解析器と用語認識器を適用することで意味構造付き MEDLINE を作成した。その意味構造を検索対象とすることで、関係概念が高精度かつ高速に検索できることを示した。

本システムでは主語-動詞-目的語の関係のみを扱ったが、意味構造はその他の関係も含んでおり、様々な概念、例えば実験環境や様相なども検索対象にできると考えられる。また、本システムで利用した基盤技術は、生物医学分野に特化したものではない。従って、関係概念の検索を必要とする様々な分野、例えば特許文書の検索や質問応答システムに応用可能と期待される。

- Miyao, Yusuke, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya and Jun'ichi Tsujii. **Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases.** In the Proceedings of COLING-ACL 2006. Sydney, Australia, pp. 1017-1024, July 2006.
- Miyao, Yusuke and Jun'ichi Tsujii. **Probabilistic disambiguation models for wide-coverage HPSG parsing.** In the Proceedings of ACL 2005. Ann Arbor, Michigan, pp. 83-90, June 2005.
- Tsuruoka, Yoshimasa 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.

- Masuda, Katsuya, 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. Edmonton, Canada, pp. 58-60, May 2003.
- Taura, Kenjiro. **GXP: An interactive shell for the grid environment.** In the Proceedings of IWIA2004. 2004.
- Ninomiya, Takashi, Yoshimasa Tsuruoka, Yusuke Miyao, Kenjiro Taura and Jun'ichi Tsujii. **Fast and Scalable HPSG Parsing.** Traitement automatique des langues (TAL). 46(2). Association pour le Traitement Automatique des Langues, 2006.

４．１．２　疾患と遺伝子の関係概念の抽出

はじめに

生命科学分野において、特定の疾患とその疾患に関与する遺伝子の関係を認識することは非常に重要な課題である。また、この分野においては学術論文のアブストラクトが網羅的に MEDLINE というデータベースに収録されている。そこで我々は、論文のアブストラクト中に記述されている特定の疾患名と遺伝子名を自動的に認識し、さらにそれらの関係を認識して、主題を分類することを目指す。

この分野において、主要な概念については既にいくつかのデータベースに知識が蓄積されているので、疾患名や遺伝子名の用語認識結果には、既存のデータベースの ID 情報を同時に付与する事によって、認識結果の有用性を高めることができる。我々は疾患名や遺伝子名の辞書を用いた用語認識の手法を用いた。

本研究では、最大エントロピー法を用いた用語認識器と関係認識器を開発し、これらをコーパスに基づいた教師あり学習に適用した。まず、MEDLINE から前立腺癌および胃癌に関連するアブストラクトを収集し、自動認識した疾患名と遺伝子名が共起する文章に対して、生物学者が用語認識の正誤と文章の主題に対する分類を付与したコーパスを作成した。この注釈付きコーパスを用語認識器と関係認識器の学習に用いた。

また、疾患と遺伝子の関係については、これらの用語が共起している文章の主題を分析することによって、関係を分類することができる。本研究においては、前立腺癌または胃癌と遺伝子の関係（DGA）を対象として、文章の主題に対して「病因」、「臨床マーカー」に関連する分類項目を仮定した実験を行った。

本節ではこの DGA 抽出システムについて詳述する。

システムの概要

本システムの概要を図１に示す。本システムでは、MEDLINE から収集した前立腺癌または胃癌に関連するアブストラクトに対して、辞書参照に基づく専門用語認識の手法を適応し、該当する疾患名と遺伝子名が共起する文を集める。この文に、機械学習に基づく用語認識器、DGA 抽出器および主題分類器を適応することによって、用語の誤認識を除去し、疾患と遺伝子の関係について記述している文を抽出し、その主題毎に分類するものである。

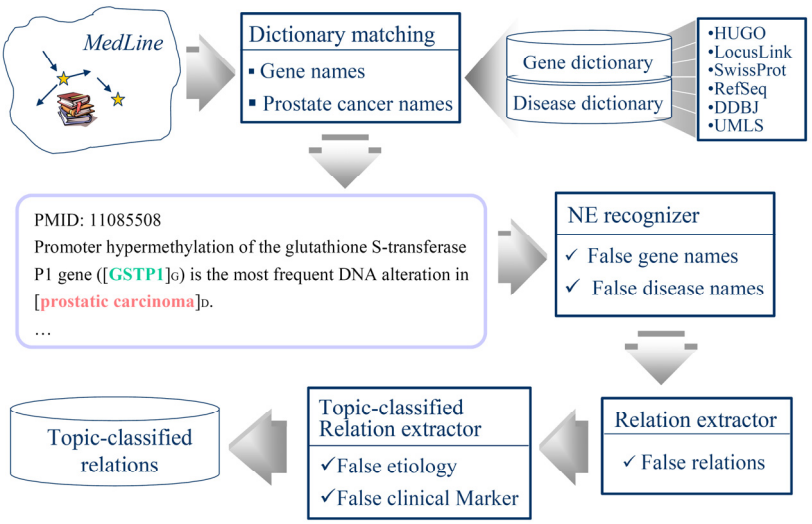


図１：DGA 抽出システムの概要

図２に主題分類の例を示す。本システムは、疾患名と関連する遺伝子のリスト、DGA の主題およびその裏付けとなる文を出力するものである。

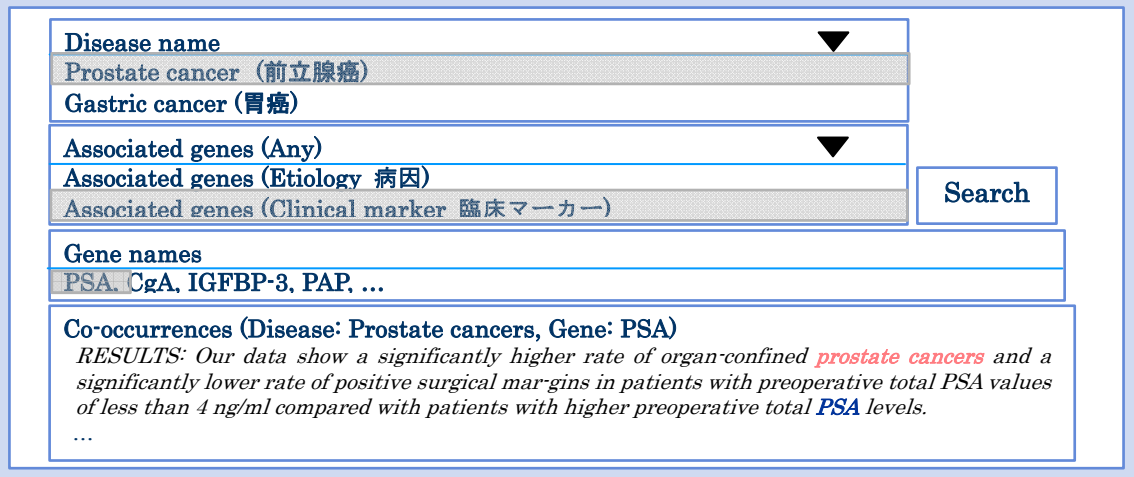


図２：文の主題分類の例

用語辞書とコーパスの作成

専門用語認識を行うための疾患名および遺伝子名の辞書は、UMLS、HUGO、LocusLink、SwissProt、RefSeq、DDBJ 等のデータベースから収集し、人手により整備したものを用いた。この辞書を用いて文中に出現する専門用語候補箇所に用語タグを付与し、疾患名と遺伝子名が共起する文を収集した。この収集の際、一文中に疾患名または遺伝子名が二回以上出現する場合には、それぞれ一つずつの組み合わせ毎に文をコピーし、疾患･遺伝子共起集合とした。この共起集合を以降のシステムの入力とする。

我々はさらに、この中から前立腺癌に関連する共起を 2,999 文、胃癌に関連するものを 1,000 文、ランダムに選出し、6 人の生物学者によって用語認識の正誤および、DGA の有無、主題の分類について正解を付与したコーパスを作成した。このコーパスを以後に述べる機械学習に基づく手法の訓練および評価データとした。

専門用語認識

前項に述べた辞書参照に基づく用語認識では、その再現率を維持するために、辞書にある文字列に一致した箇所を全て用語候補箇所としている。しかし、特に二文字や三文字の短い略語などの場合、その意味の曖昧性は決して低くない。そこで我々は、さらに機械学習による用語認識器を適応し、その精度の改善を試みた。その際、HPSG 構文解析器 ENJU(Miyao et al. 2005)と GENIA 品詞タガー(GENIA. 2004)の出力を素性とし、最大エントロピー法に基づいて機械学習を行った。この専門用語認識の効果は、後述する DGA 抽出および主題分類の精度を改善させることができた。(表１、２)

DGA 抽出

文中に共起する疾患名と遺伝子名はその疾患の原因や結果、治療効果等さまざまな観点で記述されている。しかし、場合によっては意味的に特に関係が記述されていない場合もある。そこで我々は、用語認識と同様に最大エントロピー法に基づく機械学習によって、何らかの関係が記述されている共起文を抽出し、その精度は F-スコアで前立腺癌に関する DGA が 95.5%、胃癌に関する DGA が 89.5%であった。

DGA の主題分類

前述の DGA 抽出により収集された、何らかの関係を記述する共起のうち、本研究では**病因**(Etiology)と**臨床マーカー**(Clinical marker)に関連する関係に着目し、主題分類の実験を行った。主題分類は関係概念の意味的な分類に当たるため、作成したコーパスにおいても該当する共起の数は少なく、データ疎の問題がある。そこで我々は、共参照認識による用語の同義語への拡張および、アブストラクト全体への文脈の拡張による用語の意味曖昧性解消(Yarowsky. 1995)を試みた。

結果と考察

表１と２に、用語認識、DGA 抽出、DGA の主題分類の実験結果を示す。最初の列に示した数字は､それぞれの主題毎の正解共起の数である｡評価実験においては､前立腺癌 2,999 共起、胃癌 1,000 共起からなる正解コーパスをそれぞれ 10 分割し、9 割を訓練データ、残りの１割を評価データとして、10 通りのデータセットについて学習を実施し、その平均精度を求める 10-fold cross validation によって評価を行った。2 列目に示した **P** は適合率、**R** は再現率、**F** は適合率と再現率の調和平均(F-スコア)を意味する。3 列目以降は左から順に、辞書参照に基づく用語候補の認識のみを用いた実験の結果(**Baseline w/o NER**)、機械学習による用語認識を適応した実験結果(**Baseline with NER**)、DGA 抽出実験結果(**RE**)、3 種類の条件化での主題分類実験結果(**TRE w/o RE**、**TRE with RE/Automatic**、**TRE with RE/Manual**)を示している。この結果が示すように、いずれの主題においても適合率に比べて再現率が低く、そのために F-スコアは決して高くない。そこで、文脈の拡張と共参照認識の手法を適応した。

表 2：前立腺癌関連コーパスを用いた実験結果

Prostate cancer (10-fold)		Baseline	Baseline	RE	TRE	TRE with RE	
(# of positives)	(%)	w/o NER	with NER		w/o RE	Automatic	Manual
Etiology	P	2.6	2.7	3.1	92.6	96.1	98.0
(77)	R	100.0	100.0	97.4	64.9	63.6	63.6
	F	5.1	5.3	6.0	76.3	76.6	77.2
Clinical marker	P	31.5	35.9	37.5	76.8	80.3	81.4
(945)	R	100.0	97.8	96.9	74.1	74.0	74.0
	F	47.9	52.5	54.1	75.4	77.0	77.5

表２：胃癌関連コーパスを用いた実験結果

Gastric cancer (10-fold)		Baseline	Baseline	RE	TRE	TRE with RE	
(# of positives)	(%)	w/o NER	with NER		w/o RE	Automatic	Manual
Etiology	P	16.3	18.3	21.2	82.3	85.4	87.5
(163)	R	100.0	96.9	95.7	65.6	64.4	64.4
	F	28.0	30.8	34.7	73.0	73.4	74.2
Clinical marker	P	7.1	8.4	8.8	70.7	74.5	74.5
(71)	R	100.0	95.8	95.8	57.7	57.7	57.7
	F	13.3	15.5	16.1	63.6	65.1	65.1

表 3 に示すように、前立腺癌および胃癌における病因と臨床マーカーいずれの主題分類についても、文脈拡張と共参照認識による精度の改善が見られた。

また、表４に示すように、構文解析による統語情報を用いる場合と用いない場合では、いずれの場合でも精度の改善が見られた。

表３：文脈拡張と共参照認識の効果

	Prostate cancer(F-measure)		Gastric cancer(F-measure)	
All features in	Etiology	Clinical marker	Etiology	Clinical marker
Instance sentence	75.4%	74.9%	71.0%	61.9%
+ Context extension	76.0%	75.1%	71.7%	61.9%
+ Coreference recognition	76.0%	75.5%	72.6%	63.0%

表４：構文解析の効果

	Prostate cancer(F-measure)		Gastric cancer(F-measure)	
	Etiology	Clinical marker	Etiology	Clinical marker
W/O Syntactic features	75.2%	74.9%	71.9%	61.5%
With Syntactic features	76.0%	75.5%	72.6%	63.0%

まとめと今後の課題

以上の結果から、次のような結論を得た。

(1) 用語認識、DGA 抽出および DGA 主題分類において、再現率を確保するための手法と誤認識を排除するための手法を組み合わせることにより、精度を改善した

(2) 関係概念は文単位で記述されることが多いが、その曖昧性を解消し、正しく主題を分類するためには、文脈の拡張や共参照の認識が有効であった

(3) 構文解析による統語情報は、DGA 主題分類において有効であった

今回提案した手法は、人手で作成した資源(辞書と注釈付きコーパス)に高度に依存しているが、これらの資源の構築には大きなコストがかかる。このコストを軽減するためには、動的学習法(Bonwell et.al. 1991)や分野適応(Daume et.al. 2006)などの手法を適応することが有効であると考えられる。また、本研究では、辞書参照に基づく用語認識手法を用い、その誤認識を排除する手法に重点を置いたが、今後は語の綴りの変化に対応する手法(Tsuruoka et.al. 2003)などを適応し、より再現率を改善していく予定である。

- Yusuke Miyao and Jun'ichi Tsujii, **Probabilistic Disambiguation Models for Wide-Coverage HPSG Parsing.**, Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), 2005: pp. 83-90.
- GENIA Part-of-Speech Tagger v0.3, http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/postagger/, 2004.
- David Yarowsky, (1995), **Unsupervised Word Sense Disambiguation Rivaling Supervised Methods.** Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), pp. 189-196.
- Charles C. Bonwell and James A Eison, **Active Learning: Creating Excitement in the Classroom.**, AEHE-ERIC Higher Education Report No.1, 1991: Washington, D.C.: Jossey-Bass. ISBN 1-87838-00-87.
- Hal Daume III and Daniel Marcu, **Domain Adaptation for Statistical Classifiers.**
- International Journal of Artificial Intelligence Research (JAIR)., 2006: Vol. 26, pp. 101-126.
- Yoshimasa Tsuruoka and Jun'ichi Tsujii, **Boosting Precision and Recall of Dictionary-Based Protein Name Recognition.**, Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine., 2003: pp. 41-48.

４．１．３　たんぱく質相互作用に関する情報抽出

はじめに

我々は、2006 年の第 2 回 BioCreative テキストマイニング・チャレンジにおいて、タンパク質同士の相互作用タスク（BC2 PPI IPS）[1] に参加した。我々のチーム識別子は、T19 BC2 PPI であった。本節の内容は、Sætre らの研究 （Sætre et.al. 2007） をまとめたものである。

システムの概要

この研究は、論文の本文からのタンパク質相互作用（PPI）抽出タスクに、昨年度までに開発したたんぱく質相互作用認識システム AKANE（Yakushiji 2006）がどのくらい適応できるのかを調べることから始まった。AKANE システムは、最近開発された文レベルの PPI システムで、AImed コーパス [8] において 57.3%の F スコアを達成した。AKANE システムを BioCreative タスクに適用するため、与えられた訓練データに前処理を施す必要があった。BioCreative の訓練データは、論文の本文それぞれに対し、相互作用をもつタンパク質名のペアの識別子のリストが示されている。これに対し、AKANE システムは AImed コーパスのようなアノテーション付きの文を入力として想定している。BioCreative で与えられた訓練データを AImed 形式のコーパスに変換するため、入力ファイル中の HTML タグをすべて除去して、テキスト部分のみを取り出した。専門用語タガー（NER）を適用して、テキスト中に含まれるたんぱく質名をタグ付けし、相互作用をもつタンパク質名ペアと文内で共起するタンパク質名ペアを訓練データとした。これらの処理は、パイプライン・アーキテクチャで実行される。AKANE システムの出力を BioCreative タスクの形式に変換するために、いくつかの後処理を適用した。この後処理には、結果のフィルタリング、ランク付け、F スコアを最大にするためのチューニングなどが含まれる。

システム構成

我々のシステムは、文区切り認識、専門用語認識、構文解析、タンパク質相互作用（PPI）抽出モジュールを、パイプライン・アーキテクチャで結合して実装されている。すべてのモジュールは機械学習の技術を採用し、訓練コーパス上で最大のパフォーマンスが得られるようにチューニングされている。また、これとは独立したモジュールが、BioCreative の訓練データを AImed の PPI 形式に変換する。このパイプライン・アーキテクチャを図１に示した。以降では、それぞれのモジュールの概要を述べる。

文区切り認識

生命・医学分野のテキスト向けの文区切り認識モジュールは、GENIA コーパス（Kim et.al. 2003）を訓練データとして、最大エントロピー法（MaxEnt）（Hara et.al. 2005）による分類器で構成されている。このモジュールは、ピリオド、コンマ、一重／二重引用符、閉じ括弧などを区切り記号として、文区切りの候補を認識する。このように得られた文区切りの候補に対して、分類器がその場所が本当に区切り場所として妥当かどうかを判断する。分類器の素性としては、区切り詞、前後の語、記号の存在、数値文字の存在、大文字の存在などを用いている。ここでは、語の変換ルール（コンマの除去、括弧の除去、小文字への変換）も行われる。この分区切りモジュールを、学習に用いなかった GENIA コーパス中の 200 アブストラクトで評価すると、99.7%の F スコアを得た。BioCreative の実際のテキストデータでは、HTML の特殊記号や、図のキャプションなどの差異があるため、文区切りの精度はこれよりも若干低くなるようである。

[1] http://biocreative.sourceforge.net/biocreative_2.html

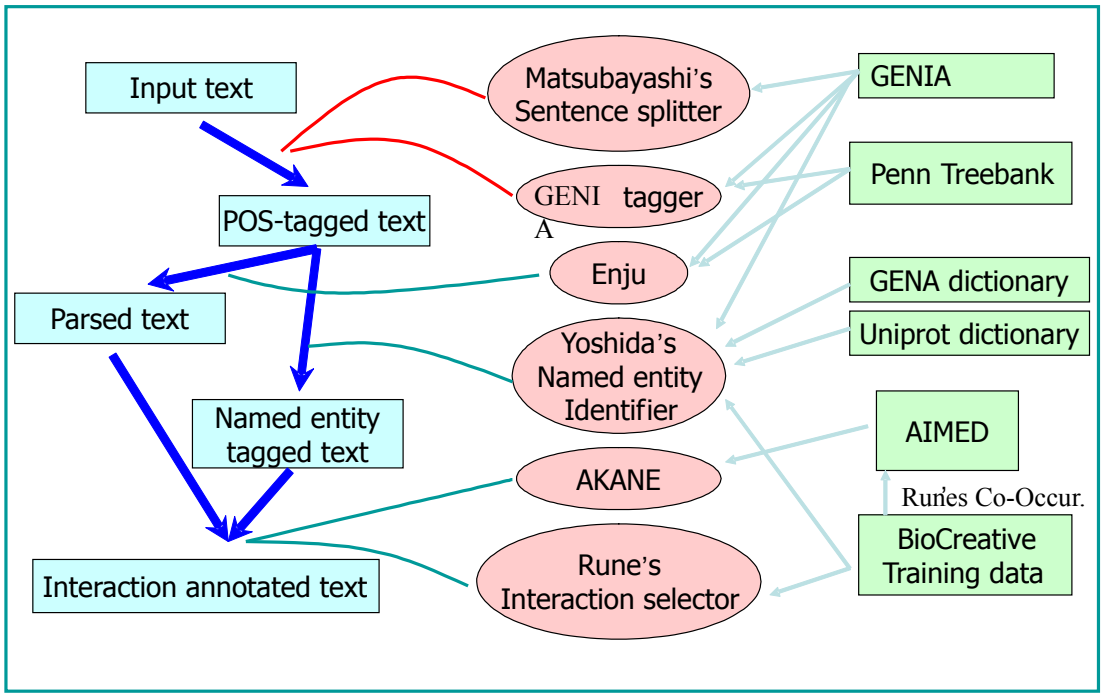


図 3：システムの概要

専門用語認識

専門用語認識モジュールは、文ごとに区切られ、品詞のタグを付与されたテキストを入力として受け取る。入力テキストを最初に処理するのは、JNLPBA（Kim et.al. 2004）共通タスクにより提供されているデータに基づいて学習した統計ベースの専門用語認識器である。この専門用語認識器は、与えられた文の部分文字列がタンパク質名である確率を出力し、ある一定の閾値以上の確率を持つ部分文字列が、タンパク質名の候補と見なされる。得られたタンパク質名の候補は、タンパク質名辞書に登録されている語と比較され、ある閾値以下の編集距離をもつ候補をタンパク質名として認識する。タンパク質名の識別子として、GENA 辞書［6］を用いて改良された Uniprot のものを採用した。どの識別子を付与すべきなのか曖昧な場合は、最大エントロピー法に基づく分類器を用いて、識別子のランク付けを行う。この分類器は、296 件の論文を学習データとし、次の素性を用いて構成されている：Uniprot エントリで示されている MEDLINE 論文と、対象とする論文の類似度；Uniprot/GENA 辞書、対象とする部分文字列と辞書エントリとの編集距離；辞書エントリのタイプ（タンパク質名、遺伝子名など）。論文同士の類似度尺度には、TF*IDF ベクトルで表現された論文のコサイン距離を用いた。それぞれの識別子には、最大エントロピー法で計算された確率が付与され、フィルタリング・モジュール（後述）に送られる。

たんぱく質相互作用情報の抽出(AKANE システム)

実際にタンパク質名のペアを抽出するのは、この AKANE システム（Yakushiji 2006）である。このモジュールは AImed コーパスを訓練データとして仮定しているので、前処理を行って、相互作用を持つタンパク質に関する文集合を得る。AKANE システムは、生命・医学分野向けの Enju HPSG 構文解析器を用いて、入力テキストを解析する。この構文解析器は Penn Treebank と呼ばれるニュース記事を用いて学習しているが、GENIA Treebank（Tateisi et.al. 2005）を用いて解析のモデルを修正するというドメイン適用手法を用い、生命・医学分野テキストにおいても高精度な解析ができる。生命・医学分野のテキストにおける性能は、F スコアで 86.9% （Hara et.al. 2005） である。AKANE システムは、専門用語認識モジュールから得られたタンパク質名ペアに関する情報と、構文解析器からの出力を統合し、両方のタンパク質名を含む最小の構文木パターンを作成する。また、複数の構文木パターンを統合し、新しいパターンを作成する。AKANE システムは、

このようにして獲得した構文木パターンのうち、どれが相互作用を正しく言い当てるのか調べるため、訓練データから構文木のパターンの頻度を計算する。AKANE システムは、可能なすべての相互作用のリストを出力するので、テキスト中で相互作用を記述している１か所に対して、複数の相互作用の候補を出力する。これは、それぞれのタンパク質名が、いくつかのタンパク質識別子として解釈できる可能性があるためである。そこで、前節で説明した専門用語認識モジュールが出力する確率を用い、最も正しいと思われるタンパク質名のペアを一つ選択する。

訓練データの生成

BioCreative コーパスに含まれる文のうち、PPI と関係するタンパク質名を 2 つ以上含む文を抽出し、AKANE システムが学習データとして受理する AImed 形式の XML ファイルに変換する。BioCreative コーパスにおいて相互作用を持つとされるタンパク質名のペアを含むすべての文は、その相互作用を記述しているものと仮定する。時間の都合により、この仮定の正確性や効果を網羅的には調べていないが、手作業で正確性を調べた範囲では、この仮定は十分妥当であると確認できた。AKANE システムは、AImed よりも規模の大きい訓練データを扱えるように設計されていないため、BioCreative コーパスに含まれる 740 件の論文のうち、250 件しか訓練に用いていない。また、専門用語認識で得られたタンパク質名／識別子を含む論文だけを訓練データとして採用したいということも、論文を絞り込んだ理由のひとつである。たくさんのタンパク質名の共起を含む論文も、AKANE システムの処理速度の問題から、除外してある。

結果と考察

16 チームが BioCreative チャレンジに参加し、45 の実験結果が BioCreative に提出された。表 1 は、システムの F スコア、適合率、再現率を、我々が提出した 3 つのシステム（Run1, Run2, Run3）、BioCreative で上位 3 位に入ったグループ、グループ全体の平均それぞれについて示したものである。より詳細な評価結果は 2007 年 4 月 22 日にスペインで開催されるワークショップにおいて発表される予定である。我々の 3 つの実験結果は、次のように作成された。Run1 は種を超えたタンパク質の相互作用を除外したもので、10.5%の F スコア（8.2%の適合率、14.6%の再現率）を示した。Run2 は、本節で説明したシステムで、13.7%の F スコア（10.6%の適合率、19.1%の再現率）を達成した。Run3 は元々の AKANE システムを AImed コーパスで訓練したものである。元々の AKANE システムは、機械学習を用いた場合 57.3%の F スコアが報告されているが、今回は時間の都合で、手作業でパラメータをチューニングし、AImed コーパスで 42.0%の F スコア（70%の適合率、30%の再現率）のシステムを用いた。それでも Run3 のシステムの性能が一番良かった。これは、BioCreative チャレンジに向けて学習コーパスを自動生成したため、学習コーパスにノイズが混ざり、生成した学習コーパスの質が、AImed コーパスの質よりも悪かったことが原因として考えられる。このように、Run3 のシステムの性能が最も良かったので、訓練データの準備方法について検討を進める予定である。この方向に向かういくつかの試みは、すでに始められている（GENIA コーパスの節を参照）。

表 3：実験結果

Name/Value	Run1	Run2	Run3	Average all groups	Top 3 groups
F-score	10.5	13.7	15.8	08.4	14.0
Precision	08.2%	10.6%	15.7%	10.2%	19.5%
Recall	14.6%	19.1%	15.9%	11.5%	19.1%

- Hara T., Miyao Y., and Tsujii J., **Adapting a probabilistic disambiguation model of an HPSG parser to a new domain**, in R. Dale, K.F. Wong, J. Su, and O.Y. Kwong, eds., IJCNLP 2005 , volume 3651 of LNAI , 199-210, Springer-Verlag, Jeju Island, Korea, October 2005, ISSN 0302- 9743.

- Kim J.D., Ohta T., Tateishi Y., and Tsujii J., **GENIA corpus - a semantically annotated corpus for bio-textmining**, Bioinformatics, 19(suppl. 1):i180-i182, 2003, ISSN 1367-4803.
- Kim J.D., Ohta T., Tsuruoka Y., Tateishi Y., and Collier N., Introduction to the bio-entity recognition task at JNLPBA, in Proceedings of the JNLPBA-04, 70-75, Geneva, Switzerland, 2004.
- Tateisi, Yuka, Akane Yakushiji, Tomoko Ohta and Jun'ichi Tsujii. **Syntax Annotation for the GENIA corpus.** In the Proceedings of the IJCNLP 2005, Companion volume. Jeju Island, Korea, pp. 222--227, October 2005.
- Sætre R., Yoshida K., Yakushiji A., Miyao Y., Matsubayashi Y., and Ohta T., **Akane system: Protein-protein interaction pairs in biocreative2 challenge**, PPI-IPS subtask, in Proceedings of the Second BioCreative Challenge Evaluation Workshop, CNIO, Spain, April 2007, ISBN 84-933255- 6-2.
- Yakushiji A., Relation Information Extraction Using Deep Syntactic Analysis, Ph.D. thesis, University of Tokyo, 2006.

４．２　“基礎技術”

４．２．１　HPSG 構文解析器の高速化

はじめに

HPSG などの語彙化文法を用いた深い統語解析は、情報検索・質問応答・機械翻訳など文の意味構造を処理対象とする高度な自然言語処理アプリケーションを実現するための重要な基礎技術である。一方、実用的なアプリケーション・システムを構築するためには、膨大なテキストデータをあらかじめ処理することが必要である場合が多い。HPSG 文法では深い解析を行うために複雑なデータ構造（**型付素性構造**）が用いられており、これに起因する解析処理の非効率性が HPSG 構文解析を実用化する際の問題点のひとつであった。

本研究では **Supertagging**、**CFG-filtering**、および**決定的曖昧性解消**という３つの技術を相補的に組み合わせることで、従来の HPSG 構文解析器に比べ約 10 倍の高速化を達成した。HPSG 構文解析は、

1. 入力文の各単語に対して語彙項目を割り当てる
2. 割当てられた語彙項目を組み合わせ、入力文の構文および意味構造の計算を行う

という２つの段階におおまかに分けられる。これら２つの段階の処理には、それぞれ

1. ある単語にたいして可能な複数の語彙項目のうち、どれをその単語に割り当てるか（**語彙的曖昧性**）
2. ある語彙項目割当てのもとで可能ないくつかの構文構造のうちどれを選ぶか（**組合せ曖昧性**）

という曖昧性がある。Supertagging は各単語に割り当てる語彙項目をごく少数に絞ることで組合せ曖昧性解消のための処理コストを減らし、構文解析全体を高速化するという既存技術である。我々は Supertagging に CFG-filtering という別の既存技術を組み合せ、より強力な語彙的曖昧性解消を、構文解析の前処理として行うことを提案した。これにより、組合せ曖昧性解消のための方法として、従来法よりもはるかに単純な、決定的曖昧性解消という方法を用いることが可能となった。評価実験の結果から、提案手法によって、型付素性構造の処理に起因する時間コストを最低限におさえることができ、構文解析処理全体の速度が既存手法に比べ 10 倍程度向上することが明らかになった。以下では、提案手法について詳述し、評価実験の結果について報告する。

HPSG 構文解析システム

図１に、提案手法の概要を示す。提案手法では、入力文に対してまず Supertagging と CFG-filtering を組み合わせた前処理を行い、語彙項目割当ての組合せ（supertag sequence）のうち大域的な一貫性のあるものを、Supertagging の際に各語彙項目に対して付与したスコアの和が大きい順に列挙する。これらの supertag sequence は順に決定的曖昧性解消器に入力され、最初に得られた構文木をシステム全体の出力とする。以下、システムの各部分について述べる。

Supertagging

語彙化文法では、単語特有の性質を語彙項目というデータとして記述する。一般に、ひとつの単語は複数の用法（例：自動詞、他動詞など）に対応する複数の語彙項目をもつ。Supertagging（Bangalore and Joshi, 1999）は入力文の各単語に対し、その単語の入力文での用法に対応する語彙項目を、周辺の単語やその品詞などの情報をもとに予測する処理である。Supertagging によって各単語に割り当てる語彙項目をごく少数に絞ることで、後続の構文解析処理が高速になる。Supertagging は単純かつ高速な処理である反面、ごく限られた情報から語彙項目を予測するため、予測された語彙項目割当てが大域的な一貫性を持たない場合がある（例：目的語となる名詞がない動詞に、他動詞を表す語彙項目を割り当てる、など）。このような語彙項目割当てが後続の構文解析器に入力として送られた場合、構文解析器は適格な構文木を見つけることができず、構文解析が失敗する。従来法ではこ

の問題を避けるために、構文木が見つからなかった場合は単語に割り当てる語彙項目の数を増やし、ふたたび解析を行うという処理を行っていた。

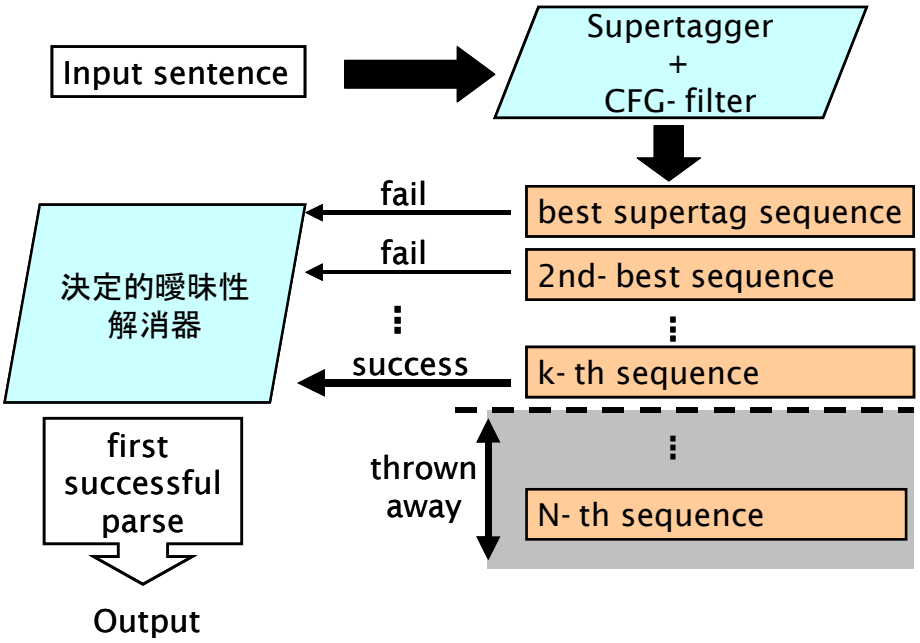


図１：構文解析システム全体

CFG-filtering

CFG-filtering（Torisawa et al, 2000）とは、HPSG 文法を近似する CFG 文法を作成し、近似 CFG による解析結果を HPSG 文法での構文解析の際に利用することで高速な解析を行う手法である。我々はこの技術を Supertagging と組み合わせ、Supertagging の結果、各単語に対し少数に絞られた語彙項目割当てから、大域的に一貫性のある組合せだけを高速に列挙するアルゴリズムを開発した。一般に近似 CFG は元の HPSG 文法と等価ではないため、選ばれた語彙項目の組合せに対して適格な構文木が存在しない場合がありうるが、実験の結果そのような例はごく少数であることが明らかになった。

決定的曖昧性解消

Supertagging と CFG-filtering を組み合わせた前処理を行った場合、構文解析器への入力は語彙的曖昧性が完全に解消された状態、すなわち各単語に対し語彙項目が一つだけ割り当てられた状態になっている。このような入力に対して、残る組合せ曖昧性を解消し、適切な構文木を見つけようとする場合、従来法よりもはるかに単純・高速な手法を用いることができる。具体的には、我々はシフト・リデュース構文解析アルゴリズムに基づく構文解析手法を用いた。この手法では、組合せ曖昧性は機械学習による分類器を用いて決定的に解消され、動的計画法を用いた従来法のように複数の部分解析結果を作成・保持しない。これにより、構文木作成の際に必要な型付素性構造の操作の回数が大幅に減少し、高速な動作が可能になる。

表１：既存手法との比較

構文解析器	40 語以下			
	LP (%)	LR (%)	F1 (%)	平均時間(ms)
提案手法	87.10	86.91	87.01	15
(Ninomiya et al., 2006)	87.66	86.53	87.09	155
(Miyao and Tsujii, 2005)	85.33	84.83	85.08	509

構文解析器	100 語以下			
	LP (%)	LR (%)	F1 (%)	平均時間(ms)
提案手法	86.90	86.71	86.80	19
(Ninomiya et al., 2006)	87.35	86.29	86.81	183
(Miyao and Tsujii, 2005)	84.96	84.25	84.60	674

表 2：CFG-filtering の効果

CFG-filter	LP (%)	LR (%)	F1 (%)	成功率(%)	平均時間(ms)
なし	88.33	84.09	86.15	79.56	556
あり	86.72	86.21	86.46	97.69	18

表 3：処理時間の内訳

サブモジュール	平均時間	全体に占める割合
POS タギング	2.7ms	17%
Supertagging	1.4ms	9%
Supertag sequence の列挙	3.8ms	23%
決定的曖昧性解消	5.2ms	32%
その他	3.1ms	19%
合計	16.3ms	100%

評価実験

標準的なデータである Penn Treebank を用いて評価実験を行った。使用した HPSG 文法は Enju 英語文法（Miyao et al., 2005）である。表 1 に、既存手法との比較実験の結果を示す。全ての結果は、同一の HPSG 文法およびテストデータに対する数値である。表中、LP、LR、F1 はいずれも解析結果の精度を表す評価値で、平均時間は一文あたりの平均処理時間である。比較対象とした構文解析器のうち、（Ninomiya et al., 2006）は Supertagging を動的計画法に基づく構文解析器と組み合わせた手法、(Miyao and Tsujii, 2005）は Supertagging を用いない、動的計画法のみを用いた手法である。結果から、提案手法は既存手法に比べ同程度の精度を保ちながら約 10 倍の高速化を達成していることがわかる。

表 2 に、提案手法において CFG-filter を使用した場合としない場合の比較結果を示す。表中、成功率は全テスト文に対する構文木が得られたテスト文の割合である。結果から、CFG-filter の使用によって成功率が大幅に増加し、処理時間は大幅に減少することがわかる。

表 3 に、1 文あたりの平均処理時間のうち、各サブモジュールで消費された時間の内訳を示す。結果より、もっとも処理時間を消費するサブモジュールは決定的曖昧性解消の部分（３２％）であることが分かる。この部分の処理時間のうち大部分は型付素性構造の操作に使われており、解析アルゴリズムの性質上、素性構造処理の回数は既にほぼ構文木を構築するための必要最低限の回数に近い。よって、例えばさらに 3 倍の高速化を実現しようとするのは非常に困難であり、逆に言えば、提案手法によって HPSG 構文解析の高速化は限界近くまで達成できたといえる。

まとめと今後の課題

本研究では、Supertagging と CFG-filtering という２つの前処理を組み合わせ、高速な HPSG 構文解析を行う手法を開発した。評価実験により、提案手法は既存手法とほぼ同程度の解析精度を保ちながら約 10 倍高速であることを明らかにした。

今後の課題としては、解析速度を維持しつつ、さらに高精度な解析を行う手法の開発が挙げられる。

- Miyao Yusuke and Tsujii Jun'ichi, **Probabilistic disambiguation models for wide-coverage HPSG parsing**, In the Proceedings of ACL 2005, Ann Arbor, Michigan, pp. 83-90, June 2005.
- Ninomiya Takashi, Matsuzaki Takuya, Tsuruoka Yoshimasa, Miyao Yusuke, and Tsujii Jun'ichi, **Extremely lexicalized models for accurate and fast HPSG parsing**, In the Proceedings of EMNLP 2006, Sydney, Australia, pp. 155-163, July 2006.
- Matsuzaki, Takuya, Yusuke Miyao and Jun'ichi Tsujii. **Efficient HPSG Parsing with Supertagging and CFG-filtering**, In the Proceedings of the Twentieth International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 1671-1676, January 2007.
- Miyao Yusuke, Ninomiya Takashi, and Tsujii Jun'ichi, **Corpus-oriented grammar development for acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank**, In the Proceedings of IJCNLP 2004, Hainan Island, China, pp.684-693, 2004.
- Kentaro Torisawa, Kenji Nishida, Miyao Yusuke and Tsujii Jun'ichi, **An HPSG parser with CFG filtering** , Natural Language Engineering, 6(1), pp. 63-80, 2000.
- Srinivas Bangalore and Aravind K. Joshi, **Supertagging: An Approach to Alomost Parsing**, Computational Linguistics, 25(2), pp. 237-265, 1999.

４．２．２　依存構造解析を利用した HPSG 構文解析

はじめに

近年、二つの点で構文解析技術が急速に発展している。一つは、CFG 構文木や依存構造などの「浅い」構文構造を対象とした構文解析で、解析精度および速度が飛躍的に向上している。これらの手法の利点は、単純なデータ構造を用いるため最先端の機械学習手法が比較的容易に適用でき、高速かつ高精度のパーザを開発しやすいことである。一方で、HPSG 構文解析に代表される「深い」構文構造を対象とする枠組みにおいては、頑健性および解析速度の飛躍的な向上により、実用アプリケーションで利用されはじめている。深い構文解析の利点は、依存構造などの浅い構造では表現できない長距離依存関係や意味構造を出力できることである。

本節では、これら二つの構文解析技術を統合することにより、HPSG 構文解析の精度を向上させる手法について報告する。本研究では、まず HPSG 構文木における主辞-非主辞関係を依存構造ととらえ、依存構造解析器を構築する。次に、依存構造解析器の出力結果を HPSG 構文解析で利用することで、HPSG 解析の精度向上を目指す。本手法は、最先端の依存構造解析技術を HPSG 構文解析に容易に利用できるという利点がある。

本研究により、HPSG 構文解析の観点からは、高速かつ高精度な依存構造解析の最先端技術を利用することで、精度向上が期待される。また、依存構造解析の観点からは、後処理として HPSG 構文解析を行うことで、長距離依存関係などのより深い構文構造や意味構造を得ることができるという利点がある。このように、本研究は、構文解析の二つの流れの両面から見て大きな意義があると考えられる。

依存構造解析

依存構造解析は、近年高速化および高精度化がさかんに研究されており、飛躍的な発展を遂げている。特に、機械学習に基づく分類器を利用した決定的依存構造解析（Nivre et al. 2004; Sagae et al. 2005）は、greedy algorithm に基づく単純な手法であるが、世界最高水準の解析精度および速度を達成している。本研究では、Sagae et al. (2005) の手法を基礎に、分類器としてサポートベクタマシンを適用した依存構造解析アルゴリズムを用いた。

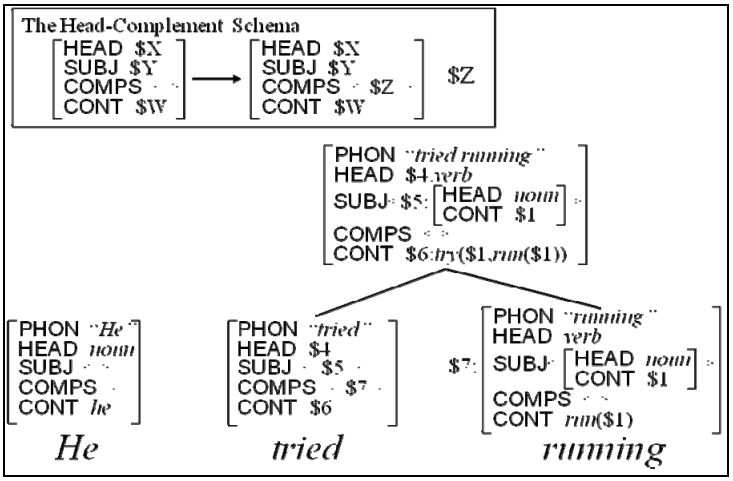


図 4：HPSG 構文解析

HPSG 構文解析

HPSG は近代言語学において代表的な文法理論であり、言語学および言語処理分野において盛んに研究されている。HPSG 理論は語彙化文法理論の一種で、文の構文構造を語彙項目とスキーマ（文法規則）で説明する。図１は、”He tried running” の構文解析過程を示している。まず構文木の終端ノードに当たる各単語に対して語彙項目を割り当てる。語彙項目は各単語の文法的・意味的制約を表現している。例えば、“running”の辞書項目は、品詞が動詞で(HEAD)、主語に名詞句を一つ取り(SUBJ)、目的語は取らない(COMPS)こと

を示している。また、CONT は意味構造を表しており、“running”の場合、意味構造が“run($1)” という述語項構造であることを示している。ここで “$1” は変数であり、これが SUBJ の CONT と同じ値であることから、構文上の主語が意味上の主語となることを表している。既存研究では、HPSG 構文解析の曖昧性解消モデルは素性森モデル(Miyao et al. 2002; 2005) および extremely lexicalized モデル(Ninomiya et al. 2006)に基づいている。これらのモデルでは、文 W に対する構文木 T の確率 $p(T|W)$を以下のように定義する。

$$p(T|W) = p(T|L,W)p(L|W) = \frac{1}{Z}\exp\left(\sum_i \lambda_i f_i(T)\right)\prod_j p(l_j|W).$$

ここで、$f_i(T)$ は構文木 T の特徴を表す関数（素性関数）、λ_i は素性 f_i の重み、$L = \langle l_1, l_2, \cdots, l_n \rangle$は各単語に割り当てる辞書項目、$Z$は正規化項である。即ち、構文木 Tの確率は、各単語に辞書項目を割り当てる確率の積と、構文木の素性の重み λ_i の積で定義される。素性関数は、適用された文法規則、句の長さ、主辞の単語など、構文木のもっともらしさを特徴づけるものを用いる。素性の重み λ_i は、最大エントロピー法により、HPSG ツリーバンクの尤度を最大化するように推定される。

依存構造解析を利用した HPSG 構文解析手法

本研究では、最先端の依存構造解析手法を HPSG 構文解析の前処理として利用することで、HPSG 構文解析の精度向上を目指す。上述の通り、依存構造解析においては比較的単純なモデルで高精度が達成できるが、HPSG 構文解析のモデルはより複雑で開発が困難になっている。HPSG は言語理論に基づく深い構文構造を計算するため、それを利用したより高度な確率モデルの構築も期待されるが、本研究では、依存構造解析と HPSG 構文解析の違いは漸進的なものであると考える。すなわち、浅い構造である依存構造はより深い HPSG 構文構造に反映されており、高精度な依存構造を与えることは HPSG 構文構造の計算にとって有用であると期待する。

特に、HPSG は語彙化理論であるため、その構文構造では単語間の依存関係が強く意識されており、これを依存構造と考えることができる。例えば、図 1 の例では、HPSG に基づく構造では、“tried”の主語と“running”の主語が共に“he”であることが変数$1 により表現されている。すなわち、“he”は同時に二つの単語に依存していることになり、これは依存構造解析の仮定、すなわち依存構造が木構造であるということに反している。しかし、“running”の主語が“he”であるという長距離依存関係を無視すれば、この文の構造は “he → tried”, “running → tried” という依存構造木で表現でき、既存の依存構造解析手法が適用できることになる。

HPSG 構文木に対して依存構造解析を適用するため、本研究では、Penn Treebank を変換することで開発された HPSG ツリーバンク(Miyao et al. 2004)を依存構造木に変換することで学習データを構築した。まず、HPSG 構文木における長距離依存関係を無視することで、語彙化 CFG 木に変換した。そして、句構造を依存構造に変換するアルゴリズムを適用することで、語彙化 CFG 木を依存構造木に変換した。これを Sagae et al. (2005)のアルゴリズムに学習データとして与え、依存構造解析器を構築した。

依存構造を HPSG 構文解析に利用する手法としては、上述の確率モデルを拡張したものを用いた。各スキーマ適用において、スキーマによって定義される主辞-非主辞の関係が、与えられた依存構造と矛盾しないかどうかを検査する。このとき、もしスキーマ適用が依存構造と矛盾するならば、対数確率値に負のスコア α を加算する。すなわち、HPSG 構文木が与えられた依存構造と n 箇所で矛盾する場合、対数確率値が $n\alpha$ 低い値となり、そのような HPSG 構文木は出力されにくくなる。

この拡張は単純であるため実装が簡単であるという利点がある。さらに、構文解析中に動的にスコアを計算するため、ビームサーチと相性が良いという利点もある。すなわち、与えられた依存構造と矛盾する部分解析木はビームサーチによって捨てられる可能性が高いため、構文解析速度を落とすことなく精度向上が期待される。

評価実験

本実験では、依存構造解析を利用した HPSG 構文解析の精度を測定するため、HPSG ツリーバンクを正解データとして述語項関係の適合率・再現率を測定した。述語項関係は $<\sigma, w_h, a, w_a>$ の四つ組で定義される。ここで、σ は述語の型（自動詞、他動詞、など）、w_h，w_a は述語と項の単語、a は関係のラベル(MODARG, ARG1, ..., ARG5)である。述語項関係の精度は、既存研究において HPSG 構文解析の精度評価で用いられている指標である。以下の実験では、Penn Treebank Wall Street Journal の Section 22 に相当する HPSG ツリーバンクを開発用、Section 23 を最終評価用として用いた。

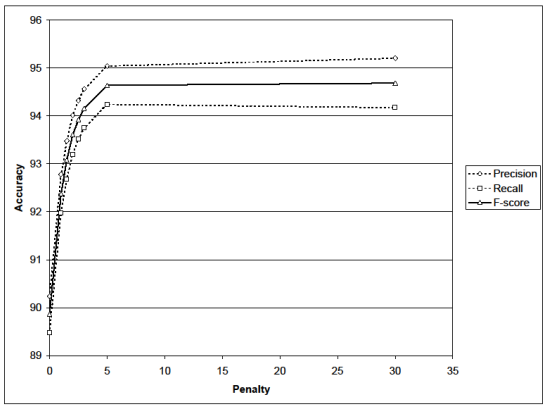


図 5：正解依存構造を与えた時の精度

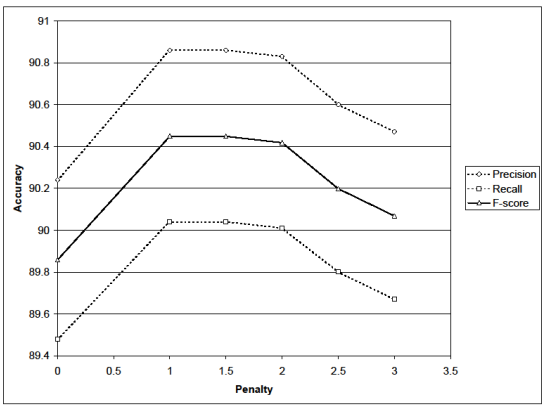


図 6：依存構造解析の出力を与えた時の精度

まず、パラメータ α と精度の関係を示す。この実験には開発用セット(Section 22)を用いた。図 2 は正解の依存構造を与えたときの精度、図 3 は依存構造解析器の出力を与えたときの精度である。図 2 から、表層的な依存構造が精度向上に大きく貢献することが分かる。また、正解の依存構造を与えたときは、α が大きいほど精度が高くなるが、α=5 でほぼ頭打ちとなっていることが分かる。一方、依存構造解析器の出力を与えたときは、α=1.5 で最高精度となり、それ以上の α では逆に精度が下がることが分かった。これは、最先端の依存構造解析器であっても間違った依存関係が含まれているため、それを過信すると精度が下がるということを示している。

次に、依存構造解析器を複数組み合わせることでより高精度な依存構造を HPSG 構文解析器に与えたときの効果を表 1 に示す。依存構造解析器を 2 つ組み合わせるときは、両解析器の出力が同じである依存関係のみを出力し、3 つ組み合わせるときは多数決により出力を決定した。2 つの組み合わせでは、8%の語についてその依存先が出力されなかったが、出力された依存関係は高精度であることが分かる。表 1 より、より高精度な依存構造を与えることにより HPSG 構文解析の精度がさらに向上することが分かる。また、利用する依存構造解析器の数が 2 つと 3 つの場合とで精度に差がなかったことから、本研究における依存構造解析では、再現率よりも適合率を重視することが必要であると言える。

パラメータ α を固定して最終評価セットで精度測定を行った結果を表 2 に示す。また、同じ評価セットを用いた既存研究の精度も示す。実験結果より、ベースラインより 1 ポイントの精度向上、既存研究より 1.4 ポイントの精度向上を達成することが示された。これらの結果は HPSG 構文解析において現在最高精度であり、本手法の有効性が示された。

まとめと今後の課題

本研究では、最先端の依存構造解析器の出力を HPSG 構文解析で利用することによって、HPSG 構文解析の精度を向上させる手法を提案した。評価実験では、ベースラインより 1 ポイントの精度向上を達成し、本手法の有効性を示した。本研究では HPSG に基づく構文解析に焦点を当てたが、提案手法は一般的なものであり、LTAG, LFG, CCG など他の語彙化文法理論に基づく構文解析でも同様の手法が適用可能と考えられる。また、本研究は、

外部の言語処理技術を導入することで構文解析の精度を向上することが可能であることを示したとも言える。依存構造解析以外にも、固有表現認識や、語義曖昧性解消などの言語処理技術が同様に構文解析に利用できる可能性が考えられる。

表 4：依存構造解析器の精度

利用した依存構造解析器の数	依存関係精度	α	HPSG	差
0	−	−	86.5	−
1	91.2	1.5	87.1	0.6
2	96.8	2.5	87.4	0.9
3	92.4	2.5	87.4	0.9

表 5：最終評価セットにおける精度

パーザ	適合率	再現率	F スコア
ベースライン	87.4	87.0	87.2
依存構造解析器 1 つ + HPSG	88.2	87.7	87.9
依存構造解析器 2 つ + HPSG	88.5	88.0	88.2
依存構造解析器 3 つ + HPSG	88.4	87.9	88.1
Miyao et al. (2005)	87.1	85.5	86.3
Ninomiya et al. (2006)	87.4	86.3	86.8

- Ninomiya, Takashi, Yoshimasa Tsuruoka, Yusuke Miyao, Kenjiro Taura and Jun'ichi Tsujii. **Fast and Scalable HPSG Parsing.** Traitement automatique des langues (TAL). 46(2). Association pour le Traitement Automatique des Langues, 2006.
- Ninomiya, Takashi, Takuya Matsuzaki, Yoshimasa Tsuruoka, Yusuke Miyao and Jun'ichi Tsujii. **Extremely Lexicalized Models for Accurate and Fast HPSG Parsing.** In the Proc. of EMNLP 2006. Sydney, Australia, pp. 155--163, July 2006.
- Miyao, Yusuke and Jun'ichi Tsujii. **Probabilistic disambiguation models for wide-coverage HPSG parsing.** In the Proceedings of ACL 2005. Ann Arbor, Michigan, pp. 83-90, June 2005.
- Miyao, Yusuke, 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. Hainan Island, China, pp. 684-693, Springer-Verlag, 2005. ISSN 0302-9743.
- Miyao, Yusuke and Jun'ichi Tsujii. **Maximum Entropy Estimation for Feature Forests.** In the Proceedings of Human Language Technology Conference (HLT 2002). March 2002.
- Nivre, Joakim and Mario Scholz. **Deterministic dependency parsing of English text**. In Proceedings of the 20th International Conference on Computational Linguistics, pages 64-70. Geneva, Switzerland. 2004.
- Sagae, Kenji and Alon Lavie. **Parser combination by reparsing.** In Proceedings of the 2006 Meeting of the North American ACL. 2006.
- Sagae, Kenji and Alon Lavie. **A classifier-based parser with linear run-time complexity.** In Proceedings of the ninth International Workshop on Parsing Technologies. 2005.

４．２．３　HPSG 構文解析器の分野適応

概要

本研究は Penn Treebank(Marcus et al. 1994)で訓練を行った HPSG 構文解析器(Ninomiya et al. 2006)を他の分野に適応する手法を提案する。本手法は、対象とする分野における単語への語彙項目割り当て確率モデルを訓練し、そのモデルを元の構文解析器に組み込む。本手法の再訓練のコストは一から曖昧性解消モデルを学習するコストに比べるとはるかに低い。

実験の結果、語彙項目の確率モデルを単純に再訓練するだけで、先行研究(Hara et al. 2005)の適応手法よりも高い構文解析精度を得ることができることが明らかになった。さらに、本手法と既存の適応手法と組み合わせることにより、一から元の構文解析器を再訓練した場合と同レベルの精度を、より低い訓練コストで達成できることが示された。

近年、語彙情報が語彙化文法ベースの構文解析器の精度に非常に重要な役割を果たすことが示されてきた(Ninomiya et al. 2006)。本研究は基本的にはこれらの研究成果に続くもので、その貢献するところは、分野のバリエーション、語彙項目割り当て確率、訓練データサイズ、および訓練コスト間の関係を実験結果として明らかにすることである。特に、本研究は、十分な精度で構文解析を行うために必要とされる対象分野のコーパス量を実験的に示す。

本報告では、最初に HPSG 構文解析器の概要と分野適応の既存手法を紹介する。次に、我々の手法である、語彙曖昧性解消モデルを再訓練し元のモデルに組み込む手法について解説する。その後、GENIA の treebank 上で我々の手法の有益性を検証する。

HPSG 構文解析器

HPSG(Pollard et al. 1994)は語彙化文法形式化に基づいた統語理論である。この理論では、少量の文法規則を用いて一般的な生成規則を記述し、大量の語彙項目を用いて各単語特有の性質を記述する。文の構造はこの 2 要素を組み合わせた木構造で表される。

本研究では HPSG 構文解析器として、Enju(Miyao et al. 2005)に一部手を加えた(Ninomiya et al. 2006)を用いる。Enju の語彙項目は Wall Street Journal の文章からなる Penn Treebank から抽出したもの(Miyao et al. 2004)で、Enju の曖昧性解消モデルは同 treebank 上で訓練されたものである。

Enju の曖昧性解消モデルは素性森モデル(Miyao et al. 2002)に基づく。与えられた文 $w = \langle w_1,\ldots,w_u \rangle$ に対して、構文解析木 t を与える確率 $p_E(t \mid \mathbf{w})$ は以下のように定義される。

$$p_E(t \mid \mathbf{w}) = \frac{1}{Z_{\mathbf{w}}} \prod_{w_i \in \mathbf{w}} p_{lex}(l_i \mid w_i) \cdot q_{syn}(t \mid \mathbf{l}),$$

$$Z_{\mathbf{w}} = \sum_{t \in T(\mathbf{w})} \prod_{w_i \in \mathbf{w}} p_{lex}(l_i \mid w_i) \cdot q_{syn}(t \mid \mathbf{l})$$

$\mathbf{l} = \langle l_1,\ldots,l_u \rangle$ は $\mathbf{w}$ に割り当てられる語彙項目のリスト、$p_{lex}(l \mid w)$ は語彙項目 l が単語 w に割り当てられる確率、$q_{syn}(t \mid \mathbf{l})$ は木構築の確率であり語彙項目のリスト $\mathbf{l}$ から構文解析木 t が得られる確率、$T(\mathbf{w})$ は $\mathbf{w}$ に割り当てられうる構文解析木の集合、をそれぞれ表す。p_{lex} および q_{syn} は、対象分野の treebank 上でその対数尤度が最大になるよう決定される。

先行研究(Hara et al. 2005)は Penn Treebank で訓練された HPSG 構文解析器を生医学分野に適用する手法を提案している。この手法では、木構築の曖昧性解消モデル q_{syn} を対象分野で再訓練することを行った。具体的には、対象分野の treebank 上で新たな q_{syn} を訓練する際に、元の構文解析器の q_{syn} を参照分布(Jelinek 1998)として取り入れた。この手法では対象分野の少量の treebank のみを用いるため、低いコストで構文解析精度を向上させることに成功した。

語彙項目割り当てモデルの再訓練

本研究での分野適応は、対象分野上で語彙項目割り当てモデルを訓練し、それを元の構

文解析器に組み込むという方策をとる。Enju は語彙項目割り当てモデルを $p_{lex}(l \mid w)$ という形で含んでいるため、上記の方策は以下のようにして実現することができる。まず適応対象分野において、新たな語彙項目割り当てモデル $p'_{lex}(l \mid w)$ を訓練し、その上で Enju 中の $p_{lex}(l \mid w)$ をこの新たなモデルで置き換える。

本研究では $p'_{lex}(l \mid w)$ として曖昧性モデル $p_{lex\text{-}mix}(l \mid w)$ を用いる。 $p_{lex\text{-}mix}(l \mid w)$ は最大エントロピーモデルで素性関数は $p_{lex}(l \mid w)$ と等しい。本手法はこれらの素性関数を用い、対象分野と元の分野の treebank を混合した訓練データ上で $p_{lex\text{-}mix}(l \mid w)$ を学習する。

実験では上記に示した本手法を実装しその構文解析精度への寄与を検証する。その際、直感的手法・既存手法も実装し、本手法と性能を比較した。実装手法は以下の通りである。

ベースライン： 元の Enju モデルをそのまま用いる
再学習（GENIA 上）： GENIA Treebank 上で Enju と同様のモデル訓練を行う
再学習（混合コーパス上）： Penn Treebank と GENIA Treebank の混合コーパス上で Enju と同様のモデル訓練を行う
既存手法： 先行研究(Hara et al. 2005)で提案された手法
本手法： 元のモデルにおける p_{lex} を $p_{lex\text{-}mix}(l \mid w)$ と置き換える（ただし q_{syn} はそのまま）
本手法＋既存手法： 元のモデルにおける p_{lex} を $p_{lex\text{-}mix}(l \mid w)$ と置き換え、さらに q_{syn} を先行研究(Hara et al. 2005)で提案された手法で再訓練する（本手法と既存手法の組み合わせ）

GENIA コーパスでの実験

上記で示された手法を実装し、その性能を比較・評価した。元の構文解析器である Enju は Penn Treebank 02-21 節（39,832 文）で開発されたもの(Ninomiya et al. 2006)を用い、再訓練の対象として GENIA Treebank(Kim et al. 2003)を選んだ。これは MEDLINE から抽出された 1,200 アブストラクト（10,848 文）からなり、本実験ではこれらを訓練用 900 アブストラクト（8,127 文）、開発テスト用 150 アブストラクト（1,361 文）、テスト用 150 アブストラクト（1,360 文）に分割して用いた。

本実験における構文解析精度は、先行研究(Ninomiya et al. 2006)と同じく predicate-argument 依存関係（ラベル適合率／再現率）の精度で与えた。各手法の性能の評価は、コストに対する精度に注目し、訓練用 GENIA treebank のサイズ、および訓練時間の変化に伴う精度変化を測定した。この際、訓練用 GENIA treebank のサイズは $100n$ (n = 1…9)アブストラクトの 9 段階で変化させた。図１と図２は訓練セットサイズ、および訓練時間の変化による精度変化を、各手法間で比較したものであり、表１は訓練セットサイズが 900 アブストラクトである場合の構文解析精度、および訓練時間を比較している。図２は「再学習（混合コーパス上)」の結果を含んでいないが、これは表１でも示されているように、他の手法に比べてこの手法のみが非常に訓練時間がかかったためである。

モデルを再訓練しない場合、Enju の構文解析は GENIA treebank に対し F-score で 86.39 の精度を与えた。これは元の分野に対する構文解析精度と比較すると 3.42 ポイント低い。これを本実験のベースラインとする。一方、「再学習（混合コーパス上)」は訓練セットがいかなるサイズでも全手法中最高レベルの精度をあげている。この手法と少なくとも同程度の精度を、可能な限り低いコストで達成することが目標となる。

本手法は、いかなるサイズの訓練用 treebank を用いてもベースラインおよび「再学習（GENIA 上)」、さらに既存手法よりも高い精度をあげている。既存手法との比較結果は、語彙項目割り当てのモデルを再訓練する方が、木構築のモデルを再訓練するよりも分野適応には重要性があることを示していると考えられる。また、少量の訓練用 treebank を用いる場合に限って言えば、本手法は「再学習（混合コーパス上で)」に近い構文解析精度を、はるかに低い訓練コストで達成していることがわかる。本手法は少量の treebank のみを用いる場合には十分な性能を期待できるアプローチと言えるであろう。

「本手法＋既存手法」はいかなる訓練データサイズでも、「再学習（混合コーパス上)」と同程度の構文解析精度をあげている。特に、最大サイズの訓練データに注目した場合、「本手法＋既存手法」は「再学習（混合コーパス上)」よりも若干高い構文解析精度を与え

ていることがわかる。この差は p 検定において有意水準 0.10 での重要性が示されており、「本手法＋既存手法」の有益な効果を期待させる結果となっている。さらに、図２と表１を見ると、「本手法＋既存手法」は「再学習（混合コーパス上)」よりもはるかに短時間しか要さないことがわかる。以上の観察により、「本手法＋既存手法」が比較した手法の中で最も望ましい手法であるということが言えるであろう。図１から、「本手法＋既存手法」は 6,500 文程度を用意すれば、元の構文解析器が元の分野に対して与えた精度と同程度の構文解析精度（表１の「ベースライン（PTB 上の精度)」参照）を達成できることがわかる。

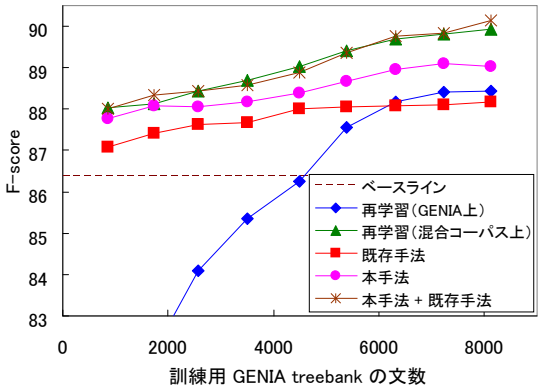


図 7：訓練データサイズと構文解析精度

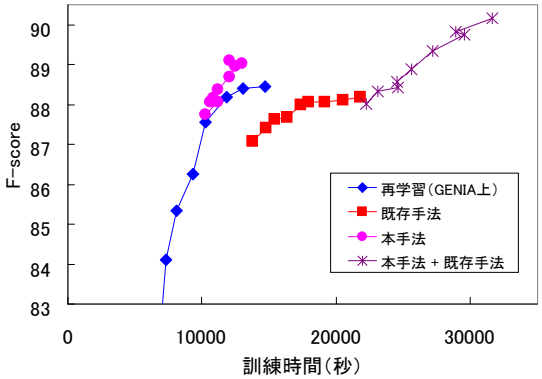


図 8：訓練時間と構文解析精度

表 6：各手法の構文解析精度と訓練時間

	F-score	訓練時間（秒）
ベースライン（PTB上の精度）	89.81	0
ベースライン	86.39	0
再学習（GENIA上）	88.45	14,695
再学習（混合コーパス上）	89.94	238,576
既存手法	88.18	21,833
本手法	89.04	12,957
本手法＋既存手法	90.15	31,637

まとめと今後の課題

本研究では Penn Treebank 上で訓練された HPSG 構文解析器を他の分野へ適応させるための手法を与えた。本手法では対象分野で語彙項目割り当ての確率モデルを訓練し、それを元の構文解析器に組み込んだ。実験では、本手法が構造的なモデルのみを再学習する既存手法よりも高い構文解析精度を与えることが可能なことを示した。さらに、本手法と既存手法を組み合わせることで、構文解析器を対象分野上で一から再訓練したモデルと同程度の構文解析精度を、はるかに低いコストで得られることを示した。本手法は 8,127 文の対象分野の treebank を用いることで構文解析精度を F-score で 3.84 ポイント向上させることができるが、6,500 文あれば、少なくとも元の分野に対するベースラインの構文解析精度を達成できることが示された。

今後の研究としては、本手法のさらなる性能向上を試みたい。例えば、分野に特化した素性すなわち、named entity のような要素を組み入れることで、対象分野の専門用語を適切に扱った構文解析を可能にしたい。

- Hara Tadayoshi, Miyao Yusuke, and Tsujii Jun'ichi, **Adapting a probabilistic disambiguation model of an HPSG parser to a new domain**, In the Proceedings of IJCNLP 2005, Jeju Island, Korea, pp. 199-210, October 2005.
- Jelinek Frederick, **Statistical Methods for Speech Recognition**, The MIT Press, 1998.
- Kim Jing-Dong, Ohta Tomoko, Tateishi Yuka, and Tsujii Jun'ichi, **GENIA corpus – a semantically annotated corpus for bio-textmining**, Journal of Bioinformatics, 19(suppl. 1):i180-i182, 2003.
- Marcus Mitchel, Kim Grace, Marcinkiewicz Mary Ann., MacIntyre Robert, Bies Ann, Ferguson Mark, Katz Karen, and Schasberger Britta, **The Penn Treebank: Annotating predicate argument structure**, In the Proceedings of the Human Language Technology Workshop, San Francisco, CA, 1994.,
- Miyao Yusuke, Ninomiya Takashi, and Tsujii Jun'ichi, **Corpus-oriented grammar development for acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank**, In the Proceedings of IJCNLP 2004, Hainan Island, China, pp.684-693, 2004.
- Miyao Yusuke and Tsujii Jun'ichi, **Maximum entropy estimation for feature forests**, In the Proceedings of HLT 2002, March 2002.
- Miyao Yusuke and Tsujii Jun'ichi, **Probabilistic disambiguation models for wide-coverage HPSG parsing**, In the Proceedings of ACL 2005, Ann Arbor, Michigan, pp. 83-90, June 2005.
- Ninomiya Takashi, Matsuzaki Takuya, Tsuruoka Yoshimasa, Miyao Yusuke, and Tsujii Jun'ichi, **Extremely lexicalized models for accurate and fast HPSG parsing**, In the Proceedings of EMNLP 2006, Sydney, Australia, pp. 155-163, July 2006.
- Pollard Carl and Sag Ivan A., **Head-Driven Phrase Structure Grammar**, University of Chicago Press, 1994.

４．２．４　品詞タガーの分野適応

はじめに

品詞タグ付けは自然言語処理における基本的かつ重要な処理のひとつである。品詞タグは、単語の基本形を求める処理やパターンベースの情報抽出処理などに直接利用されるだけでなく、固有表現抽出や構文解析の入力としても利用される (Okanohara et al., 2006; Yoshida et al., 2007)。そのため、処理の対象とする文書に対して、精度の高い品詞タガーを構築することは、実用的な言語処理を実現する上できわめて重要である。

近年の機械学習を利用した品詞タグ付けに関する研究によって、適切なアルゴリズムと大量の訓練データを用いれば、標準的な英語コーパスにおいて 97%を超える精度が実現できることが示されている。最高精度を実現しているアルゴリズムとしては、Cyclic Dependency Networks (Toutanova et al., 2003), Support Vector Machines (Gimenez and Lluis, 2004), Bidirectional Maximum Entropy (Tsuruoka and Tsujii, 2005) などが挙げられる。

機械学習ベースの手法の最大の問題は、学習に必要な訓練データを作成するためのコストである。例えば、バイオ分野のコーパスである GENIA コーパスでは、約２万の文に対して、品詞情報が付与されているが、それにかかる人的・時間的コストはきわめて大きい。新たな分野に言語処理を適用しようとするたびにそれだけのアノテーションのコストが必要とされるのでは、実用的な言語処理を実現することは難しい。そこで本節では、最小限の学習データで、高精度の品詞タガーを構築するための手法に関して述べる。

CRF による品詞タグ付け

高精度の品詞タガーを構築する上で最も重要なのは、どのようなモデルを利用して学習を行うかである。本研究では、品詞タグ付けのための学習モデルとして、Conditional Random Field (CRF) を用いる。CRF は、近年注目を集めている確率モデルのひとつであり、固有表現認識やチャンキングなどのタスクで高い性能が報告されている。CRF では、文全体に対して、識別モデルとしてひとつの対数線形モデルを定義し、学習データの尤度を最大化するようにモデルのパラメータを決定する。

従来、CRF は英語の品詞タグ付けに適用するには学習コストが大きすぎると考えられてきたが、我々の実装では、パラメータの数値最適化アルゴリズムに改良をすることで、WSJ コーパス全体を使った場合でも学習時間を約 7 時間程度に抑えることが可能となった。また、精度に関しても、Wall Street Journal (WSJ) コーパスのテストデータにおいて、ほぼ最高レベルの精度である 97.18%を達成した。

生成モデルの利用

機械学習に基づく自然言語処理においてタグ付きデータを削減する手法の一つとして、生コーパス（品詞情報が付与されていないテキスト）の利用が挙げられる。確率モデルを使っている場合、それが生成モデルであれば、EM アルゴリズムを使って生コーパスを活用することができる。しかし CRF は識別モデルであるため、生コーパスを直接利用することはできない。

Kazama (2001) らは、識別モデルにおいても生コーパスを活用する方法として、生成モデルの出力を識別モデルの素性として利用する手法を提案し、最大エントロピー法による品詞タガーを用いた実験でその効果を確認している。そこで本研究ではまず、CRF でも同様の効果があるかを確認する。

生成モデルとしては隠れマルコフモデルを用いる。すなわち、隠れマルコフモデルの学習を Baum-Welch アルゴリズムを用いて生コーパス上で行う。CRF を学習する際には、各文において、隠れマルコフモデルの出力 (Viterbi アルゴリズムで得られる隠れ状態) を、各単語に関する素性として追加する。

能動学習

教師付き学習では、通常、学習データはランダムサンプリングによって生成されるものとする。それに対して、能動学習（active learning）と呼ばれる枠組みでは、学習データのサンプリングの方法を工夫し、情報量の大きい訓練データを重点的にサンプルすることで、少ない学習データで高精度を達成することを目指す。

本研究では、能動学習の一手法である、Uncertainty Sampling と呼ばれる方法を用いて学習データのサンプリングを行う。基本的な考え方としては、タグ付けが難しい、すなわち高い信頼性でタグ付けができないような文を重点的にサンプリングして、アノテーションを行っていくという手法である。以下にそのアルゴリズムを示す。

1．コーパス全体からランダムサンプリングによって得られたｋ個の文に対してアノテーションを行い、それらを用いてタガーの学習を行う。
2．コーパスの各文に対して、現時点のタガーを利用してタグ付けを行う。その際、それぞれの文において、タグ付けがどれほど信頼できるのかを、各トークンに対する最も確からしいタグの周辺確率の平均値として定量化する。
3．コーパス全体から、最も信頼度が低くタグ付けされたｋ個の文を抽出し、それらについてアノテーションを行う。
4．アノテーション済みのすべての文を用いてタガーの学習を行う。
5．タガーの精度が満足できるレベルになるまで２に戻り繰り返す。

異分野のコーパスで訓練されたタガーの利用

新たな分野（対象ドメイン）のための品詞タガーを構築する際、場合によっては、別な分野（元ドメイン）のコーパスを活用することが可能である。たとえば、バイオ分野を対象とした品詞タガーを構築する場合、ニュース分野のコーパスである WSJ コーパスを利用することが考えられる。もっとも素直な利用方法として、元ドメインのコーパスを対象ドメインの学習データに加えるという方法が考えられる。しかしこの方法の場合、学習コストが大きくなるという問題だけでなく、タグセットが同一でなければならないことや、元ドメインのコーパスの影響が大きくなりすぎるという問題がある。

本研究では、元ドメインで学習されたタガーの出力を、対象ドメインのタガーの学習の素性として利用するというアプローチをとる。これにより、学習時には、対象ドメインのデータだけを扱えばすむため、学習のための計算コストが小さくなり、能動学習のように繰り返し学習を行うような枠組みにも対応できるようになる。

実験

各手法の有効性を確認するため、WSJ コーパス、GENIA コーパス、PennBioIE コーパスを用いて実験を行った。WSJ はニュース分野、GENIA、PennBioIE はバイオ分野のコーパスである。

図１に、WSJ コーパス上での実験結果を示す。図中の曲線は、訓練データの量と精度との関係（学習曲線）を示している。”Random”は、通常のランダムサンプリングによる学習、”Random+HMM” は、それに隠れマルコフモデル（HMM）からの素性を追加して学習を行った結果である。”Active Learning + HMM” は、さらに能動学習を行った場合の結果である。HMM の素性を加えることにより、とくに、学習データが少ない段階で精度が大きく向上していることがわかる。学習データが増えてくると、能動学習の効果が顕著になってくる。能動学習と HMM との組み合わせにより、100000 単語の学習データ（5000 文弱）で 96.71%の精度を達成している。

次に、バイオ分野のコーパス上での実験結果を示す。図 2 に、GENIA コーパスを利用して実験を行った結果を示す。”+WSJ”は、元ドメイン（WSJ コーパス）で学習したタガーの出力を素性として利用していることを示している。元ドメインのタガーからの素性を加えると、学習データが少ない段階での精度が大きく向上することがわかる。能動学習と組み合わせることにより、80000 単語（約 3000 文）の学習データで 98.40%の精度を達成している。これは GENIA コーパス全体（約２万文）を利用して学習を行った場合の精度

（98.58%）にかなり近く、本研究の学習データ削減手法の効果がきわめて大きいことがわかる。

PennBioIE コーパスでの実験結果を図 3 に示す。各手法の効果に関して GENIA コーパスとほぼ同様の傾向が見て取れる。能動学習によって、100000 単語（約 6000 文）でほぼ上限に近い 97.66%の精度を達成している。

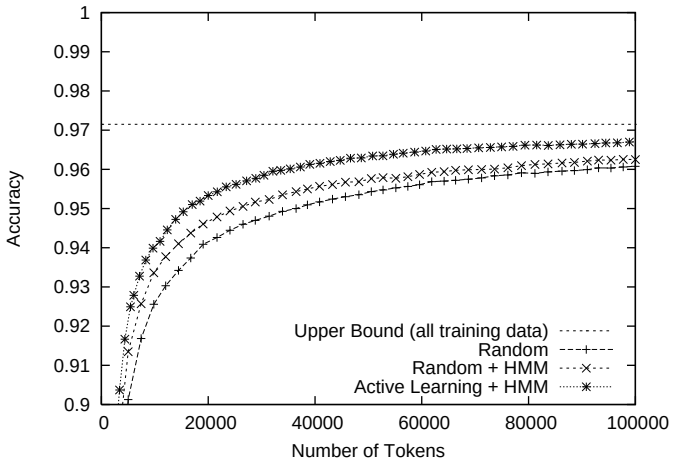


図 1　Wall Street Journal コーパスでの学習曲線

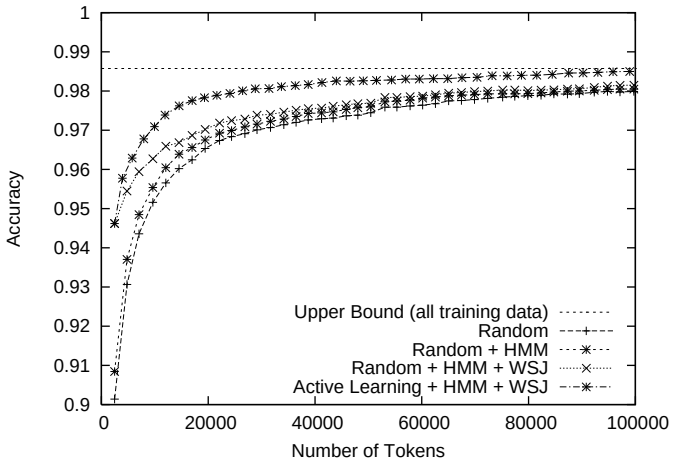


図 2　GENIA コーパスでの学習曲線

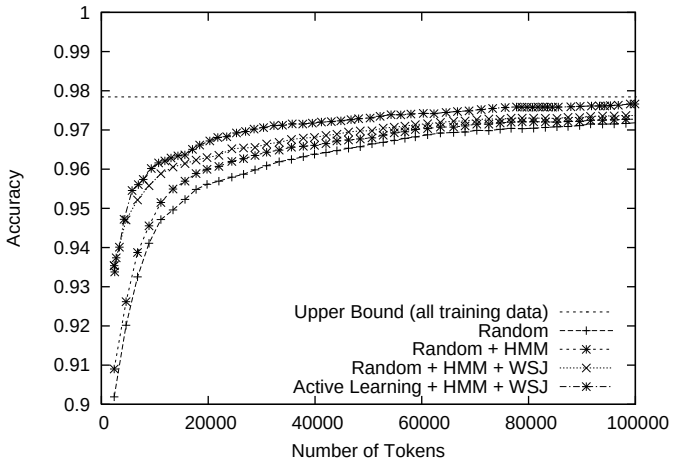


図 3　PennBioIE コーパスでの学習曲線

まとめと今後の課題

本研究では、学習データ削減のための手法（生成モデルの利用、別ドメインタガーの利用、能動学習）を組み合わせることにより、タガーを新たな分野に適応させるための学習データの量を著しく削減できることを示した。GENIA コーパスおよび PennBioIE コーパスを利用した実験では、数千文の学習データで、コーパス全体を使った場合とほぼ同レベルの精度を達成している。

本研究では、品詞タグ付けを対象タスクとして実験を行ったが、同様の手法は、固有表現認識やチャンキングなど他のタスクでも有効だと考えられる。CRF は、それらのタスクでも高い精度が報告されており、固有表現・チャンキングなどへの適用は興味深い課題である。

- Yoshida, Kazuhiro, Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Tsujii. **Ambiguous Part-of-Speech Tagging for Improving Accuracy and Domain Portability of Syntactic Parsers.** In the Proceedings of IJCAI-07, pp. 1783-1788, 2007.
- Okanohara, Daisuke, Yusuke Miyao, Yoshimasa Tsuruoka, and Jun'ichi Tsujii. **Improving the Scalability of Semi-Markov Conditional Random Fields for Named Entity Recognition.** In the Proceedings of COLING/ACL 2006, pp. 462-472, 2006.
- Tsuruoka, Yoshimasa and Jun'ichi Tsujii. **Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data.** In the Proceedings of HLT/EMNLP 2005. Vancouver, pp. 467-474, 2005.
- Kazama, Jun'ichi, Yusuke Miyao, and Jun'ichi Tsujii. **A Maximum Entropy Tagger with Unsupervised Hidden Markov Models.** In the Proceedings of NLPRS 2001, pp. 333-340, 2001.
- Toutanova, Kristina, Dan Klein, Christopher Manning and Yoram Singer. **Feature-rich part-of-speech tagging with a cyclic dependency network.** In the Proceedings of HLT-NAACL 2003, pp. 252-259, 2003.
- Gimenez, Jesus, and Lluis Marquez. **SVMTool: A general pos tagger generator based on support vector machines.** In the Proceedings of LREC 2004, 2004.

４．２．５　領域代数に基づく構造付きテキスト検索システム

はじめに

近年のテキストに対する様々なタグ付け手法の発展により、膨大なテキストに対してあらかじめ様々な情報を付与し、その情報を利用した情報抽出、情報検索が可能になってきている。テキストから情報を得る際にキーワード検索の結果をその後解析するのではなく、検索時により複雑な情報要求をシステムに対して行うことで、その後の解析を省略し、同じ時間でより複雑な解析が可能となる。

このようにテキストに対して様々な情報を付与する形式としては XML が最も一般的である。XML に対しては XQuery、XPath といった問い合わせ言語が標準化として存在しており、それらを実装した XML データベース(Meier 2002, Boncz et al. 2006)も多く存在する。しかしながら、それらの XML データベースは大規模なデータに対して適用できるものがなく、また XML の定義からタグ領域の交差を許さないものが多い。様々な方法によってテキストにタグ付けが行われた場合、タグ領域が交差しないという保障は存在しないため、交差にも適応できるデータベースが必要となる。

そこで本節では、領域代数を基にした構造化テキスト検索システムについて報告する。本システムでは、領域集合間の演算として定義される領域代数を拡張し、高速かつ大規模データに対しても適用できる構造付きテキスト検索を実現する。また、検索要求中に変数を導入することで、テキスト中で離れた位置にある部分を結びつけることができる。また、その変数を用いた検索要求に対する高速なアルゴリズムを提案する。

領域代数

領域代数は「領域集合間の演算の集合」として定義される。表 1 に演算の例を示す。

表 7：領域代数

A ▹ B	B を含む領域 A
A ◃ B	B に含まれる領域 A
A △ B	A と B を含む領域
A ▽ B	A または B を含む領域
A ◇ B	A で始まり B で終わる領域

これらの各演算に対しては同名のタグ領域間に包含関係がない(入れ子構造が存在しない)ことを前提に高速なアルゴリズムが提案されている(Clarke et al. 1995)。また、入れ子構造の存在を考慮したアルゴリズムも提案されている(Jaakkola et al. 1999)。しかしながら、後者のアルゴリズムでは演算の際にほぼ全ての領域を探索する必要があるため、大規模データベースに対しては適用できない。そこで、前者のアルゴリズムを改良し、入れ子構造が存在するデータに対しても高速に演算可能なアルゴリズムを提案する。

インデックス構造

テキスト中の各単語、タグ(開始タグ、終了タグ)およびタグ中の属性・属性値組に対してインデックスを作成する。インデックスには、位置情報および深さ情報を格納する。深さ情報により、開始タグと終了タグのマッチングが容易になり、探索する際にも深さ情報を利用して入れ子構造の親子関係が容易に知ることができる。検索に必要なインデックスを単純な構造に、高速なアルゴリズムを使用することで大規模な文書集合に対しても適用可能で高速なシステムとなる。

探索アルゴリズム

探索アルゴリズムの一部を図 1 に示す。左図は Clarke らによるアルゴリズムであり、右図は改良後のアルゴリズムである。これらは A ▹ B を満たす最初の領域を探索するアルゴリズムである。τ (Q,k)は位置 k から見て領域の開始位置基準で最初の Q を満たす領域を表し、ρ (Q,k)は位置 k から見て終了位置基準で最初の Q を満たす領域を表す。改良後のアルゴリズムでは、一度 A ▹ B を満たす A の領域を発見した後、その A の領域を包含する別の A の領域が存在するかを確認し、存在した場合には新たに発見した方の A の領域を探索

結果として出力する。同様にして他の演算子に対しても入れ子構造の存在を考慮したアルゴリズムとなるよう改良を行った。

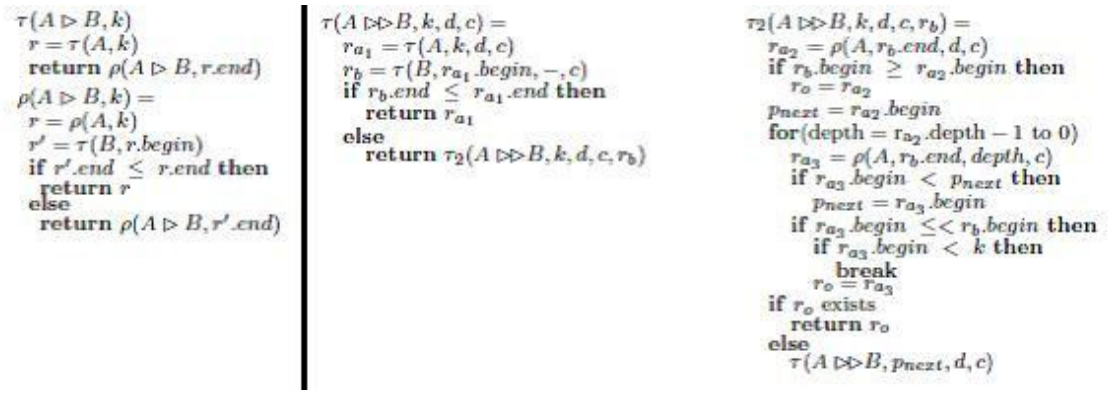


図 9：探索アルゴリズム

変数の導入

変数を使用することにより、テキストに付与された情報を利用してテキスト中で離れた位置にある部分を結びつけることができる。例えば図２の XML 文書に対して

[sentence] ▷ (([word arg1="$s"] ▷ activate) △ ([phrase id="$s"] ▷ P53))

というクエリは$s の部分に共通の値が代入され(例では$s=1)、「activate の主語は P53 である」ということを表し([name]はタグ領域を表す)、テキストに付与されているタグの属性 arg1 および id を利用することによってテキスト上では離れた位置にある”activate”と”P53”を結びつけることができる。

```
<sentence sentence_id="27426f">
 <phrase id="0" cat="S" head="4" lex_head="7">
  <phrase id="1" cat="NP" head="2" lex_head="3">
   <phrase id="2" cat="NP" head="3" lex_head="3">
    <word id="3" pos="NN" cat="NP" base="p53">p53</word>
   </phrase>
  </phrase>
  <phrase id="4" cat="VP" head="5" lex_head="7">
   <phrase id="5" cat="VP" head="6" lex_head="7">
    <phrase id="6" cat="VP" head="7" lex_head="7">
     <word id="7" pos="VBZ" cat="VP" base="activate" arg1="1" arg2="8"
     rel_type="activation">activates</word>
    </phrase>
    <phrase id="8" cat="NP" head="9" lex_head="10">
    …
```

図 10：構造つきテキスト例

変数に入る値を決定する際に、一度に全ての値を決定しようとすると複雑な処理が必要となってしまうため、クエリを細かい単位に分解し変数の値を順番に一つずつ決定し、最後に決定した値を元のクエリに代入し、実際に検索を行うことで正しいかどうかを確認する。また、順番に変数の値を決定する際にも変数間の依存関係を考慮するためにある変数に対する値が決まったら、それ以降はその変数にはそれに対する値が代入されているものとして計算を行う。

変数付きクエリに対するアルゴリズムは以下のとおりである(簡単のためにある条件を満たす[sentence]を検索すると仮定する)。

1. クエリ中の全ての単語を含む[sentence]を探索する
2. その[sentence]内で変数の値を決定する

3. 決定した変数の値を元のクエリに代入し、その[sentence]内で検索を実行する

変数値の決定は以下のアルゴリズムで行う。

1. 与えられたクエリから部分クエリを作成する
2. それらの部分クエリの中で、変数を一個含みかつその部分クエリを実行した際の検索結果が最も少ないと推定される部分クエリを選び出す
3. その部分クエリから変数部分を取り除いたクエリを実行し、検索されたテキストから変数の値を決定する
4. 決定した変数の値を元のクエリに代入する
5. 全ての変数に対して値が決定するまで１.から繰り返す

変数値計算の複雑な処理を行う前にキーワード検索により検索対象の絞込みを行うことで複雑な処理を行う対象を少なくし、効率的に検索を行う。また、変数値計算においても変数に入りうる値の候補数が少ない順に各変数の値を決定していくことで効率的に検索を行う。

評価実験

以上のアルゴリズムを実装し、実験を行った。検索対象は構文解析器 Enju による構文解析結果を XML 形式で付与された MEDLINE アブストラクトとした。データサイズ等を表２に示す。このデータに対して検索を行い、検索時間を計測した。検索に使用した検索要求を表３に示す。これらをそれぞれ表４に示す領域代数表現に変換し実行した。検索時間および件数を表５に示す。膨大なデータに対して高速に検索が行われていることが分かる。

表 8：検索対象(MEDLINE)のサイズ

文献数	14,785,094
アブストラクト数	7,291,857
文数	70,935,630
単語数	1,462,626,934
MEDLINE テキストのサイズ	9.3 GByte
構文解析結果付きテキストのサイズ	292 GByte
位置-深さインデックスのサイズ	219 GByte

表 9：検索要求

1	P53 activates WAF1
2	P53 does not activate *something*
3	*Something* activates P53
4	*Something* interacts with CD4

表 10：検索要求に対する領域代数表現

No.	Converted Query
1	'[sentence] ▷ ([word arg1= "$subject" arg2="$object" base="activate"] △ ([phrase id= "$subject"] ▷▷ ([word] ▷ p53)) △ ([phrase cat="np" id="$object"] ▷▷ ([word] ▷ WAF1)))'
2	'[sentence] ▷ (([word arg1="$subject" id="$verb" base="activate"]) △ ([phrase id="$subject"] ▷▷ ([word] ▷ p53)) △ ([word arg1="$vp0"] ▷ not) △ ([phrase id="$vp0" lex_head="$verb"]))'
3	'[sentence] ▷ ([word arg2="$object" base="activate"] △ ([phrase id="$object" cat="np"] ▷▷ ([word] ▷ p53)))'
4	'[sentence] ▷ (([word id="$verb" base="interact"] △ ([word arg1="$vp" arg2="$object"] ▷ with) ▷ [phrase id="$vp" lex_head="$verb"]) △ ([phrase cat="np" id="$object"] ▷▷ ([word] ▷ CD4)))'

表 11：検索時間、検索件数

No.	検索時間(1 件)	検索時間(全件)	検索件数
1	0.152s	1.271s	8
2	0.086s	2.265s	11
3	0.005s	4.251s	578
4	0.049s	3.971s	111

次に、既存の XML データベース eXist と比較を行った。検索対象は 133,375 件の構文

解析結果を付与された MEDLINE アブストラクトとした。インデックスの際、eXist はおよそ 1/3 にあたる 48,095 件のアブストラクトを非常に複雑であるとしてインデックスすることができなかった。本システムは各単語およびタグに位置情報と深さ情報を付与しているだけの単純なインデックスを用いているため、このような複雑さによりインデックス作成が行えないということは起こらない。

本システムと eXist による検索時間を表 6 に示す。本システムでの検索の際には表 4 で示した表現と同様の領域代数表現に変換し、eXist による検索の際には表 7 で示す XQuery 表現に変換し検索を行った。既存の XML データベースに比べて非常に高速であることが分かる。

表 12：検索時間

検索要求	本システム	eXist
P53 activates *something*	0.107s	58.806s
Something activates P53	0.104s	58.838s
P53(keyword のみ)	0.015s	11.556s

表 13：検索要求の XQuery 表現例

P53 activates *something*	for $s in //sentence[.//word = “p53”], $w in $s//word[@base=”activate”], $p in $s//phrase[.//word = “p53”] where $w//@arg2=$p/@id return $s

まとめと今後の課題

本節では、領域代数を用いた高速な構造化テキスト検索システムについて報告した。既存の高速なアルゴリズムを入れ子構造に対しても適用できるように改良し、さらに変数を導入することでより複雑な検索を可能にした。また実際のデータで既存の XML データベースとの比較実験を行い、XML データベースに比べて高速かつ頑健であることを示した。

本システムでは包含関係および and/or を指定した検索が可能であるが、より複雑な質問要求に答えるためにはさらに高度な演算が必要であると考えられる。また、本システムを利用することで大規模データベースに対して複雑な検索を行うことができ、そのような検索が必要とされる質問応答システムなどに応用可能であると期待される。

- Clarke, Charles L. A., Gordon V. Cormack, Forbes J. Burkowski **An Algebra for Structured Text Search and A Framework for its Implementation.** The Computer Journal, 38(1):43-56, 1995.
- Jaakkola, Jani and Pekka Kilpelainen **Nested Text-Region Algebra**. Technical Report C-1999-2, University of Helsinki, 1999.
- Boncz, Peter, Torsten Grust, Maurice van Keulen, Stefan Manegold, Jan Rittinger and Jens Teubner **MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine**. In the Proceedings of SIGMOD2006, pp. 479-490, 2006.
- Meier, Wolfgang **eXist: An Open Source Native XML Database**. In Web, Web-Services, and Database Systems: NODe 2002 Web- And Database-Related Workshops, 2002
- Miyao, Yusuke and Jun'ichi Tsujii. **Probabilistic disambiguation models for wide-coverage HPSG parsing.** In the Proceedings of ACL 2005. Ann Arbor, Michigan, pp. 83-90, June 2005.
- World Wide Web Consortium **XQuery 1.0: An XML Query Language**. http://www.w3.org/TR/xquery/
- World Wide Web Consortium **XML Path Language (XPath) 2.0**. http://www.w3.org/TR/xpath20/

４．３　“言語リソースの構築”

４．3．１　GENIA コーパス

はじめに

学術文献などでは情報は主に自然言語で書かれている。そのため、コンピュータは情報に直接アクセスできない。一般に、コンピュータがよりよく言語を解読することができれば言語にコード化された情報によりよくアクセスすることができると期待されている。このため、生物医学分野の文献からのテキストマイニングにおいても、自然言語処理（NLP）技術のアプリケーションがますます使われるようになっている。

ここ数十年間、コーパスは常に NLP 研究の中心にあった。コーパスとは、特定の領域またはスタイルを代表するテキストまたは言語表現の大きなコレクションであり、注釈付きコーパス（あるいはタグ付きコーパス）とは、計算可能な解釈をテキストに付けたものを指す。注釈（タグ）をつけることによって、コンピュータはテキストから抽出すべき情報と関係するテキストの部分に直接アクセスすることが可能になる。適切に設計されたタグ付きコーパス(Penn Treebank、MUC データセット、TREC データセットなど)は、NLP、情報検索（IR）または情報抽出（IE）の研究分野において、計算機システムの設計や試験のための参照データとして、研究課題の設定や研究の推進に貢献してきた。

GENIA コーパスは、MEDLINE データベースに登録された分子生物学文献アブストラクトのコレクションであり、この分野のテキストの代表例となることを目的としている。GENIA コーパスの主な価値は、いろいろなレベルでコーパスに与えられた注釈にある。それらの注釈はこの分野の文献に関する参照データとして研究コミュニティに貢献してきた。実際に、GENIA コーパスに基づく多くの研究成果が報告されている。

GENIA コーパスにおける注釈付けは、2 つの観点からなされている。一つはテキストにコード化された生物医学的な知識を明示すること（意味的注釈）であり、もう 1 つはテキストの統語構造を明らかにすること（言語的注釈）である。

言語的注釈

テキストの言語構造（例えば句構造や依存構造）はテキストマイニングを実際におこなう人々の関心の主な対象ではないかもしれないが、言語構造解析の精度向上がテキストマイニングシステム全体の精度向上につながるとして、しばしば研究対象となってきた。テキストの言語構造がわかるということは、完全なものではないにしろ、テキストに埋められた知識の鉱脈への経路や場所を示す地図を持っているようなものであり、テキストの言語構造に関する情報がテキストにコード化された情報にアクセスするのに有効であるということは一般に認められている。

文のトークン（英語では単語とトークンはほぼ一致すると考えられる）への切り分けと品詞付けは、しばしば、文の基本単位とその特性（例えば文法上あるいは統語上のアイデンティティ）を決定する、NLP 処理の最初のステップと考えられている。図 1 にトークンに切り分けられて品詞タグが付けられた文の例を示す。句読点と括弧は通常、隣接するトークン（単語）から切り離されて別のトークンとみなされることに注意する。

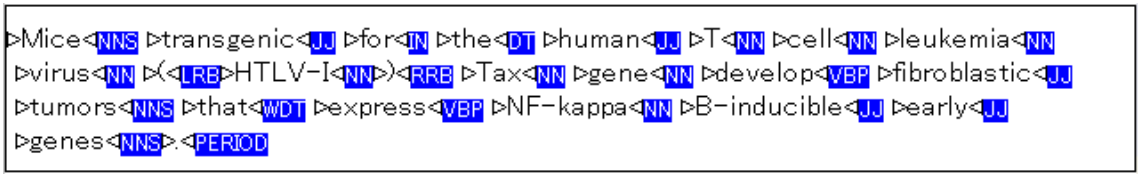


図1. 品詞タグ付きテキスト

我々は MEDLINE 上のアブストラクト 1,999 件について、デファクトスタンダードである Penn Treebank の品詞付けスキーマにもとづく 42 種類の品詞を文中の各トークンにつけたコーパスを作成した。Tsuruoka ら (2005)は品詞付き GENIA コーパスを訓練コーパ

スとして使用するとMEDLINEテキストの品詞付けの精度が91.6% から98.5% に向上するとしている。

表 1. GENIA コーパスの使用による品詞タガーの性能の向上

訓練コーパス	テストコーパス	
	WSJ	GENIA
WSJ	97.2%	91.6%
WSJ + GENIA	97.2%	98.5%

統語解析は、文中の語がどのように組み合わされて文の意味を形成するかを明らかにする。統語解析では、文中のいくつかの語がまとまって句を形成し、さらにそれらと別の語や句がまとまってより大きな句を形成する、という現象が文全体がひとまとまりになるまで繰り返される、と考えられている。そして、全部の文をカバーしている要素を根、句と一致している要素を節、語と一致している要素を葉とする木構造を与える。

図2に、統語構造をタグ付けした文を示す。

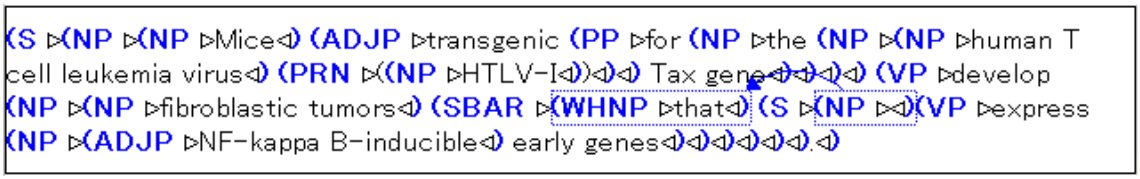


図2. 木構造タグ付きテキスト

我々はMEDLINE上のアブストラクト1,200件について木構造をタグ付けした(GENIA Treebank)。GENIA Treebankにおいても、品詞コーパス同様、デファクトスタンダードであるPenn Treebank II (PTB) ガイドライン(Beis et al, 1995) にできる限り従った。Hara ら (2005) によれば、GENIA Treebank を使用することによって HPSG パーザのF-スコアを85.1% から 86.9% にすることができる。

意味的注釈

タンパク質名や遺伝子名のような医学・生物学分野の専門用語は医学・生物学的な研究の対象となる最も基本的な構造であり、そのような用語がテキスト中のどこでどのように参照されているかを見つけることは役に立つ情報にアクセスするために重要である。GENIAでは、医学・生物学的に意味のある専門用語に、それらの意味クラスをタグとして付けている。専門用語の範囲の定義と意味クラスは、GENIA オントロジに基づく。

図3に専門用語タグが付いている文を示す。例えば、「Mice」は、多細胞生物(Multi_cell)というタグが付けられている。専門用語は再帰的（ある用語の中に他の用語が含まれる）でありうる。例えば、「human T cell leukemia virus（HTLV-1）Tax gene」には「human T cell leukemia virus」、「HTLV-1」、「Tax」の3つの専門用語が含まれ、「human T cell leukemia virus（HTLV-1）Tax gene」自身も専門用語であるとタグ付けされる。

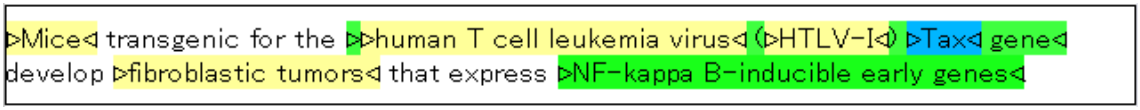


図 3. 専門用語タグ付きテキスト

GENIA オントロジは医学・生物学分野に重要な概念を定義するもので、GENIA 専門用語タグは GENIA オントロジに基づいて定義される。 図 4 に GENIA オントロジを示す。GENIA オントロジでは概念は階層的に分類される。階層の終端に当たる概念が図中では太線で囲まれているが、これらが文献上でタグ付けされるべき専門用語に付けられるタグとなる。終端概念の横に現れる数は GENIA コーパス・バージョン 3.01 でそのタグの出現し

た頻度を示す。

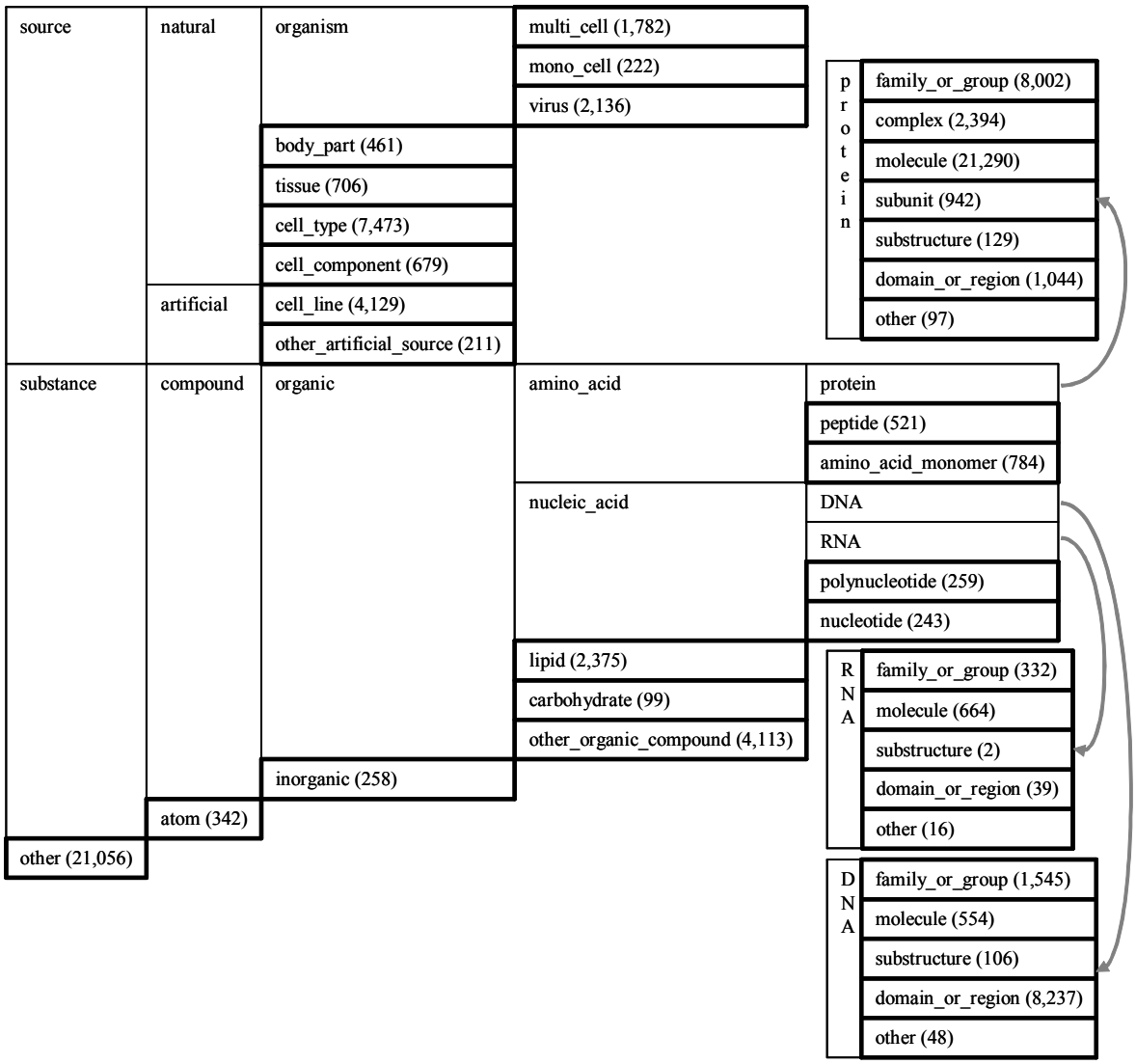


図 4. GENIA オントロジの階層構造

MEDLINE 上の 1,999 アブストラクトに対して GENIA オントロジで定義された用語タグを付与した。 GENIA 専門用語コーパスは現在デファクトスタンダードの 1 つとして広く使用されている。表 2 に GENIA 専門用語コーパスを使用して訓練された、医学･生物学分野の用語認識システムの最先端のものをいくつか示す。

表 2. GENIA コーパスを使用したエンティティ認識システムの性能

システム	再現率	適合率	F-スコア
SVM+HMM (Zhou et al., 2004)	76.0	69.4	72.6
Semi-Markov CRFs (in prep.)	72.7	70.4	71.5
Two-Phase (Kim et al., 2005)	72.8	69.7	71.2
Sliding Window (in prep.)	71.5	70.2	70.8
CRF (Settles, 2005)	72.0	69.1	70.5
MEMM (Finkel et al, 2004)	71.6	68.6	70.1
:	:	:	:

まとめと将来の課題

GENIA コーパスは品詞、構文木、生物医学分野の専門用語をタグ付けしたコーパスである。多くの情報のつけられた GENIA コーパスは、生物学分野のテキストマイニング研究者・開発者に利用され、高性能なシステムが GENIA コーパスを利用して作られた。将来の課題の一つとしては、アブストラクトのみならず論文本文にまでコーパスの対象を広げることがある。論文コーパスを整備することによって論文本文には含まれるがアブストラクトに含まれない情報を抽出するシステムの構築が可能になると期待される。

- Tateisi, Yuka 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. Lisbon, Portugal, pp. 1267-1270, May 2004.
- Tsuruoka, Yoshimasa, 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.** In the Advances in Informatics - 10th Panhellenic Conference on Informatics. LNCS 3746. Volos, Greece, pp. 382--392, November 2005. ISSN 0302-9743.
- Tateisi, Yuka, Akane Yakushiji, Tomoko Ohta and Jun'ichi Tsujii . **Syntax Annotation for the GENIA corpus.** In the Proceedings of the IJCNLP 2005, Companion volume. Jeju Island, Korea, pp. 222--227, October 2005.
- Hara, Tadayoshi, 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. Lecture Notes in Artificial Intelligence3651. Jeju Island, Korea, pp. 199--210, Springer-Verlag, October 2005. ISSN 0302-9743.
- Kim, Jin-Dong, 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. ISSN 1367-4803.
- Kim, Jin-Dong, 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). Geneva, Switzerland, pp. 70--75, 2004.
- Ohta, Tomoko, Yusuke Miyao, Takashi Ninomiya, Yoshimasa Tsuruoka, Akane Yakushiji, Katsuya Masuda, Jumpei Takeuchi, Kazuhiro Yoshida, Tadayoshi Hara, Jin-Dong Kim, Yuka Tateisi and Jun'ichi Tsujii. **An Intelligent Search Engine and GUI-based Efficient MEDLINE Search Tool Based on Deep Syntactic Parsing.** In the Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. Sydney, Australia, pp. 17--20, July 2006.

４．３．２　GENIA イベントアノテーションとその利用

はじめに

前節に述べたように、我々はこれまでに品詞、木構造、専門用語といった、言語的注釈および意味的注釈を付与した GENIA コーパスを構築し、公開してきた（Tateisi et.al. 2004, Kim et.al. 2004, Tateisi et.al. 2005）。我々はさらに深い意味的注釈として、アブストラクト中に記述される生命科学的事象への生物学者による注釈付け（アノテーション）を行い、このコーパスを用いてイベント認識の実験を試みた。本節ではこのアノテーションと認識実験について述べる。

GENIA イベントタグ付きコーパス

GENIA イベントタグ付きコーパスは、既に専門用語タグを付与した GENIA コーパス（MEDLINE アブストラクト）に対して、生命科学的事象（イベント）に関する記述に人手で注釈を付けたものである。注釈の対象となるイベントは、以下に述べるイベントオントロジーにあらかじめ定義された概念を用いてその意味クラスを与えられる。

このコーパスでは、958 件のアブストラクト（7,992 文）に対してアノテーションを行い、約 34,000 のイベントが記述されている。

図１に GENIA イベントオントロジーの階層構造を示す。ここに挙げる概念は、生命科学分野オントロジのデファクトスタンダードである *Gene Ontology* に基づいて定義した。その際、*Gene Ontology* の網羅範囲外の概念である **Artificial_process**（主に実験的な操作等）、抽象的または複合的な概念である **Correlation**（具体的に記述されていないが、互いに影響を与え合う）と **Gene_expression**（遺伝子が発現する過程を複合的に表現）の 3 つの概念を独自に定義した。

図１における各概念の横に示した数字は、GENIA コーパス中で記述されたイベントの数である。

今回アノテーションの対象とするのは、生命科学的事象すなわち、生物学的なエンティティが何らかの状態の変化を起こす過程であり、主に以下のようなイベントがある。（ただし、生命科学的に重要な情報であっても、状態の変化に関するものではない記述はアノテーションの対象ではない）

- 遺伝子の発現　（IL2 gene expression）
- 転写の開始（initiation of IL2 transcription）
- レセプターの活性化（TCR triggering）
- 細胞の分裂の促進・抑制（T cell proliferation is initiated by …）
- 複合体の形成（CD229-Grb2 complex formation）
- 触媒反応（PT catalyzed ADP-ribosylation）
- ウイルスの感染（HIV infection）

上記のようなイベントを、以下に述べる要素を用いて記述するものとする。

- EVENT　＝　生物学的な状態変化の過程の１つ１つを EVENT とし、下記の TYPE, THEME, CAUSE, CLUE を用いて記述する
- TYPE　＝　イベントの種類（イベントオントロジーから指定）
- THEME　＝　イベントの対象（term, event, sentence などの ID 指定）
- CAUSE　＝　イベントの原因（term, event, sentence などの ID 指定）
- CLUE　＝　sentence 中の手がかり表現を記述
- COMMENT　＝　イベントに対して自由記述

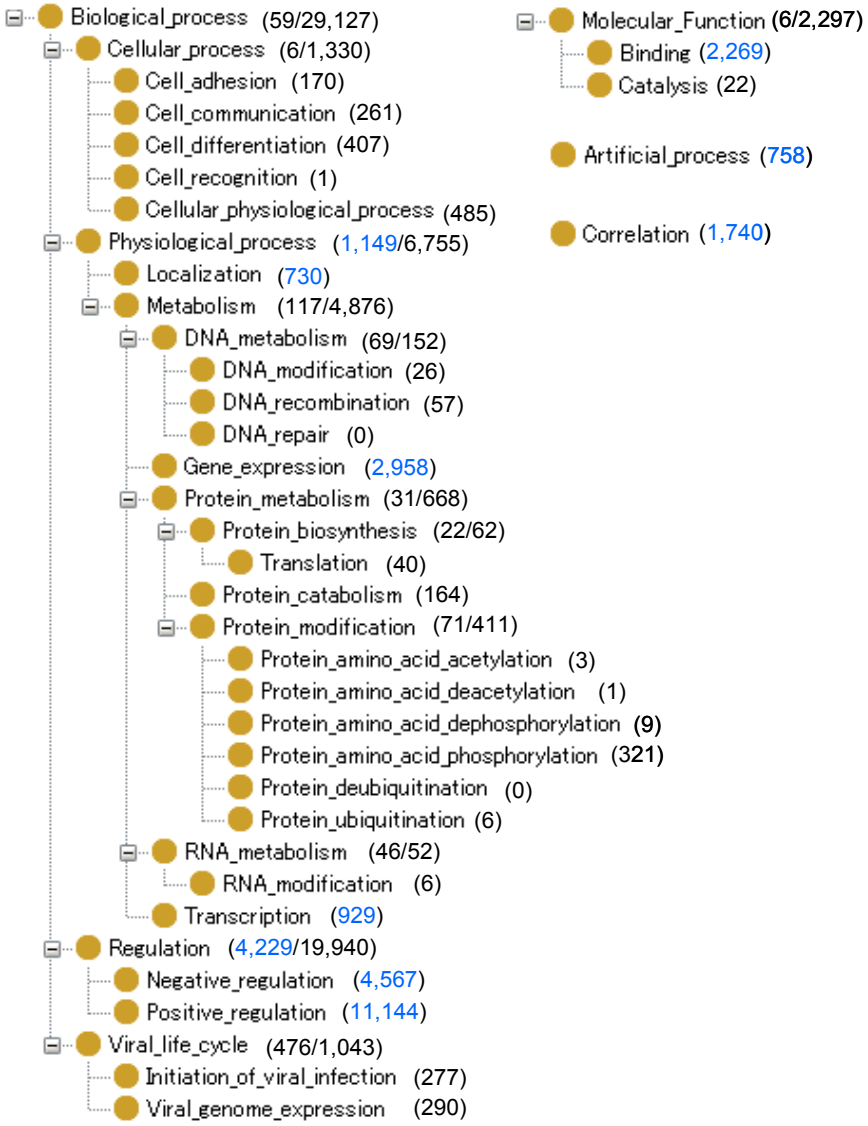


図 11：GENIA イベントオントロジーの階層構造

図２にイベントタグ付きテキストの例を示す。黒い線で囲まれたテキストが専門用語タグを付与した GENIA コーパス中の文であり、それぞれの用語はその ID と、意味クラスに対応する背景色が付与されている。この文中に表現されているイベントに関する情報を、前述の要素を用いて記述している。

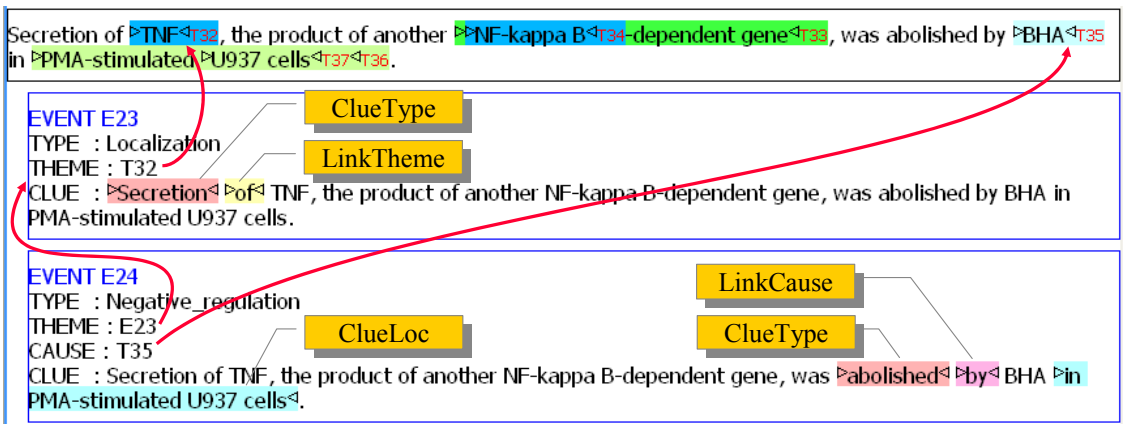


図 12：イベントタグ付きテキスト

THEME および CAUSE 要素では､既にタグ付けされている用語の ID を参照している。CLUE 要素の中では、clueType、clueExperiment、clueTime、linkCause、linkTheme、corefCause、corefTheme などのタグを使用することができ、イベントの意味クラスを決定する手がかりとなった表現や、実験・時間に関する表現、CAUSE や THEME につながる表現などに注釈をつけることができる。また、CAUSE や THEME が代名詞などを用いて表現されており、文中に対象とする用語が存在しない場合は、他の文章中に出現する ID を参照することが許されているが、その場合は参照元となる語にも注釈をつけることができる。

イベント認識実験

我々は、ある文中に特定のタイプのイベントが記述されているか否かというイベント認識の問題を、二値分類の問題として扱えることを実証するために、この GENIA イベントコーパスを用いて、*Cellular_physiological_process*（GO:0050875）タイプのイベントの認識実験を試みた。ここに述べる手法は、イベントコーパスを用いれば他のタイプのイベントにも容易に一般化することができる。我々はこの認識実験に、素性を自動的に学習するための最大エントロピー法に基づく分類器と、文の統語構造を利用するための HPSG 構文解析器 **Enju**（Miyao et.al. 2005）を用いた。

GENIA イベントコーパス 7,992 文中、*Cellular_physiological_process* タイプのイベントの注釈が付与されている文は 337 文である。本実験のベースラインの一つは、全ての文をこのタイプのイベントを記述するものであると分類するのもであり、この場合、適合率は 4.2%（下限）、再現率が 100%（上限）であることになる。

まず我々は、文中に出現する語を素性として用いて実験を行った。ここでは、文を語の集合として捉え、語順は考慮しない。この素性を用いた実験では、適合率 43.8%、再現率 40.3%であった。

次に我々は、コーパス中に記述される手がかりの表現を用いて実験を行った。図 3 はこのタイプの注釈の例であり、この文に記述されるイベントは”proliferation”という語が手がかりとなって、*Cellular_physiological_process* タイプであると記述されている。

The effects of prostaglandin E2 (PGE2) on cytokine production and **proliferation** of the CD4+ human helper T cell clone SP-B21 were investigated.	
EV1	TYPE: Cellular_physiological_process THEME: CD4+ human helper T cell clone SP-B21 CLUE_EXPRESSION: proliferation

図 3：*Cellular_physiological_process* の例

このような手がかりの表現をコーパスから収集し、語尾変化を吸収した上でキーワードとして用いた。表 1 に実験に用いた代表的なキーワードとその統計を示す。図 3 の例文で見られた”proliferation”という表現は 10 行目の”prliferat”に相当するが､このキーワードに一致した場合は 71%の確立で *Cellular_physiological_process* タイプのイベントであると期待されることになる。このキーワードリストには、非常に適合率の低いものも含まれているが、キーワードを素性として最大エントロピー法に基づく分類器を用いて実験を行った結果、適合率は 64.7%、再現率は 58.5%であった。

このキーワードを用いる実験では、周囲に出現する単語を用いてキーワードの曖昧性解消を行ったが、我々はさらに、**Enju** による構文解析の結果を分析し、イベントの記述には特徴的な構文上のパターンがあることがわかった。そこで我々は、この統語パターンを素性として用いた。表 2 に実験に用いた統語パターンを示す。パターン 1 では、イベントの THEME となる表現と、キーワードとが同じ名詞句の中に出現することを意味する。パターン 2 は THEME の表現を含む名詞句が、キーワードを含む名詞句と前置詞”of”によって導かれていることを示す。また、パターン 3 は THEME の表現を含む名詞句が、キーワードを含む動詞句の意味上の主語または目的語であることを意味している。これらの統語パ

ターンを素性として用いた実験では、適合率 65.2%、再現率 61.5%を達成することができた。

表１：イベント認識に用いたキーワード

	Keyword	precision	# appearance	# annotation
1	diapedesis	100%	3	3
2	cytolysis	100%	2	2
3	division	100%	2	2
4	rolling	100%	2	2
5	mitogenesis	85%	7	6
6	cytoly	8%	10	8
7	expansion	75%	4	3
8	lysis	72%	11	8
9	syncyti	71%	7	5
10	proliferat	71%	153	109
…				
42	divid	8%	65	5
43	heterologous	6%	31	2
44	clear	5%	40	2
45	stimulat	4%	519	23
46	form	4%	222	8

表２：イベント認識に用いた統語パターン

Pattern1	[KEYWORD THEME_ENTITY]**NP**
Examples	highly polarized TH2 clone, cell proliferation,
Pattern2	[KEYWORD]**NP** of [THEME_ENTITY]**NP**
Examples	proliferation of cell, grow of multiple mold-4 cem cell
Pattern3	[KEYWORD]**VP** —(ARG)→ [THEME_ENTITY]**NP**
Examples	proliferate human monocyte, kill fibroblastic keratinocyte-derived human cell

まとめと今後の課題

これまでに開発・公開してきた GENIA コーパスに、さらにイベントについての注釈付けを行った。現時点ではまだ、その注釈の一貫性は十分に高いとは言えないため、修正を加えている段階であり、この修正作業の後に公開する予定である。

また、このコーパスを用いて特定タイプのイベントの認識実験を行ったが、今後さらに以下のような課題が考えられる。

① 実験に使用する素性：今回の実験は、GENIA イベントコーパスを用いて行い、その精度は十分割交差検定法によって計算した。しかし、GENIA コーパスは MEDLINE 全体から見ると非常に偏った、小さな部分集合に過ぎない。MEDLINE 全体からイベントに関する記述を認識するためには、より豊富でグローバルな素性を用いていくことが必要かもしれない。

② イベントのクラス：GENIA イベントオントロジーに定義するイベントクラスは、コーパス中の文に記述されるイベントを分類するのに適度な抽象度であると考えている。しかし、注釈の一貫性確保のために行う修正作業の過程で、Gene Ontology に規定されるより上位または下位の概念を用いることが適切かもしれない。

いずれにしても、このイベント認識手法を実用化していくためには、さまざまなイベントクラスに適応できるように、頑健で一般化された分類器を開発することが重要である。

- Tateisi, Yuka 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. Lisbon, Portugal, pp. 1267-1270, May 2004.
- Tateisi, Yuka, Akane Yakushiji, Tomoko Ohta and Jun'ichi Tsujii . **Syntax Annotation for the GENIA corpus.** In the Proceedings of the IJCNLP 2005, Companion volume. Jeju Island, Korea, pp. 222--227, October 2005.
- Kim, Jin-Dong, 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. ISSN 1367-4803.
- Miyao, Yusuke and Jun'ichi Tsujii. **Probabilistic disambiguation models for wide-coverage HPSG parsing.** In the Proceedings of ACL 2005. Ann Arbor, Michigan, pp. 83-90, June 2005.

５．類似研究の国内外の研究動向・状況と本研究課題の位置づけ

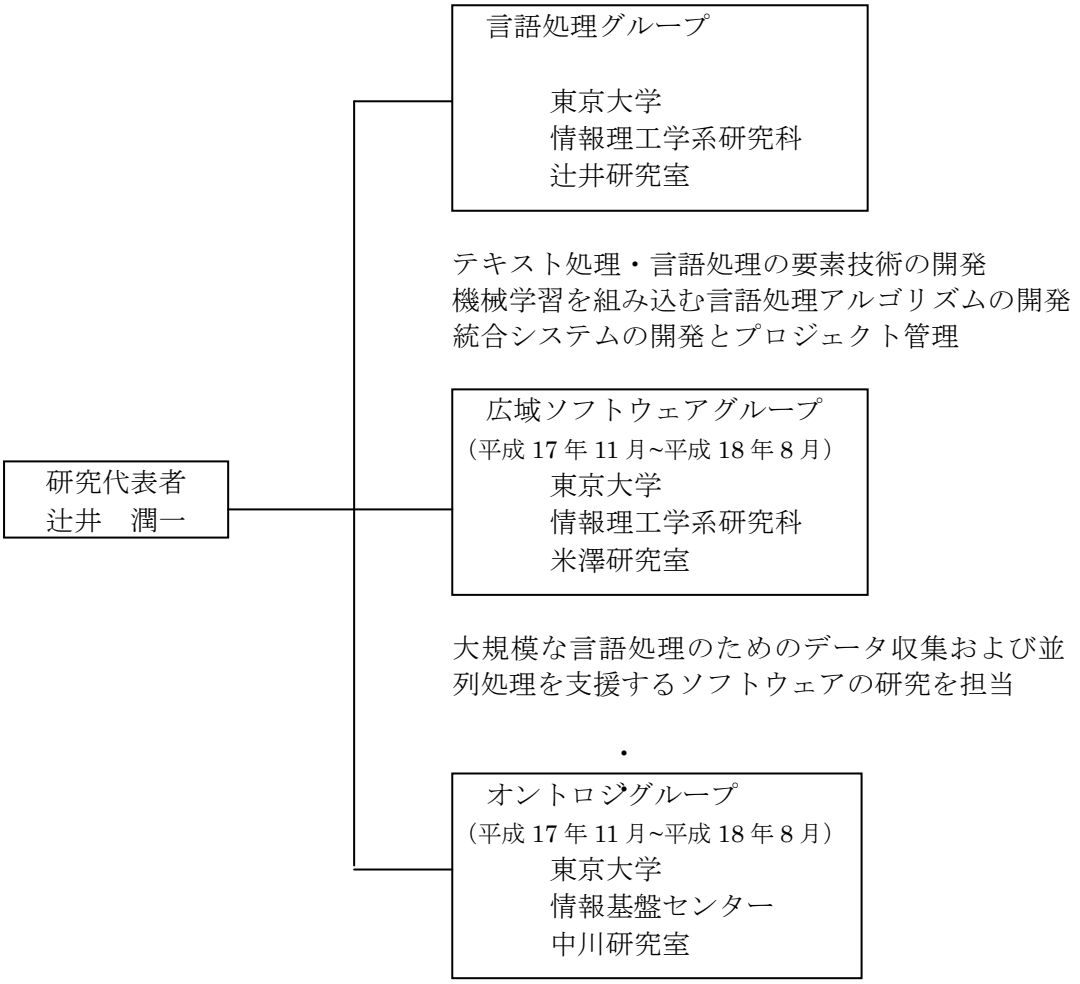
[研究の全体的な枠組み]
本研究は、本格的な言語処理技術をテキスト検索、テキストマイニングに適用する枠組みを確立することを目指し、しかも、その有効性を生命科学におけるテキスト処理に適用することで、その有効性を確認することを目指した。
　高度な言語処理の研究、構造つきデータの索引構造の研究、GRID計算環境の研究、検索に言語処理の結果を使う研究、生命科学におけるテキストマイニングの研究は、世界的に見ても、それぞれ個別に行われており、本研究のように、包括的な研究を目指したものは、ない。

［高度な言語処理］
先行する CREST で開発し、その洗練を本研究で行った HPSG に基づく英語解析器は、この分野での世界のトップグループの一つとなっている。ほかには、エディンバラ大学のCCG、Parc のLFG、ケンブリッジのHPSG、ペンシルベニア大学の LTAG の研究が、我々と誉い成果を出している。ただ、高速、高精度、分野適応の機能を備え、かつ、7000万文、1．4億語という巨大なテキストを実際に処理したシステムは、ない。
　平成18年度半ばに、Parc の R,Kaplan 博士のグループが、スピンアウトのベンチャー企業を設立し、Wikipedia の巨大テキストを LFG で処理するという計画を発表した。この試みが成功すれば、理論的な基盤の強さと巨大なテキスト量を処理できる高耐性を持ったシステムとして、我々と同等な技術をもつ唯一のグループとなる。

［生命科学分野のテキストマイニング］
この分野は、現在、急速に発展しており、言語処理をあまり使わないマイニング技術を使った商用システムも存在する。ただ、言語処理技術を使わないシステムの限界が明らかになるにつれ、我々が目指している方向の研究を行うグループが増加している。我々のグループ、エディンバラ大学、コロンビア大学、コロラド大学、ケンブリッジ大学、EBI(European Bioinformatics Institute)、スタンフォード大学などが主要なグループであるが、チューリヒ大学、シンガポール国立大学、ノルウェー工科大学、KAIST、台湾 Academia Sinica などにも、も研究グループができるなど、研究グループの数は増大している。
　言語処理応用の第一歩としてのNER研究に、我々の GENIA コーパスがゴールドスタンダードとして使われるなど、本研究の成果は世界をリードするものとなっている。このことは、MEDIE,Info-Pubmed など本研究の成果が広く注目されていること、また、本研究の研究代表者（辻井）が、英国が設立した国立テキストマイニングセンター(National Centre for Text Mining)の研究所長に任命されたことでも示されている。

[GRID 環境を用いたテキスト処理]
我々のグループが巨大なテキスト処理を日常的に行えるのは、広域ソフトウェアグループが構築したGRID環境が自由に使えるためである。現在のところ、テキスト処理にGRID環境を有効に使っているのは、日本のグループのみであり、世界をリードしている。

６．研究実施体制
(1)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
（平成 17 年 11 月~平成 18 年 8 月）
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
オントロジグループ
（平成 17 年 11 月~平成 18 年 8 月）
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

(2)メンバー表

(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	東京大学大学院情報理工学系研究科	教授	言語処理と統括	全期間
宮尾祐介	東京大学大学院情報理工学系研究科	助手	文法フォーマリズム	全期間
建石由佳	工学院大学	助教授	オントロジの作成	全期間
鶴岡慶雅	英国マンチェスター大学	研究員	機械学習	全期間
金　進東	東京大学大学院 情報理工学系研究科	研究員	情報抽出	全期間
大田朋子	東京大学大学院 情報理工学系研究科	研究員	コーパスの作成	全期間
樽川美佳	派遣先	事務員		全期間
伊藤美奈子	東京大学大学院 情報理工学系研究科	事務員		全期間
薬師寺あかね	東京大学大学院 情報理工学系研究科	博士課程	情報抽出	全期間
狩野芳伸	東京大学大学院 情報理工学系研究科	博士課程	文法学習	全期間
荒木淳子	東京大学大学院 情報学環	博士課程	自動抄録	全期間
岡崎直観	東京大学大学院 情報理工学系研究科	博士課程	情報抽出	全期間
増田勝也	東京大学大学院 情報理工学系研究科	博士課程	テキストデータベース	全期間
松崎拓也	東京大学大学院 情報理工学系研究科	博士課程	構文解析	全期間
全弘宇	東京大学大学院 情報理工学系研究科	博士課程	情報抽出	全期間
王悦	東京大学大学院 情報理工学系研究科	博士課程	オントロジの作成	全期間
綱川隆司	東京大学大学院 情報理工学系研究科	博士課程	機械翻訳	全期間
吉田和弘	東京大学大学院 情報理工学系研究科	博士課程	機械学習	全期間
大内田賢太	東京大学大学院 情報理工学系研究科	博士課程	文法フォーマリズム	全期間

原忠義	東京大学大学院 情報理工学系研究科	博士課程	文法学習	全期間
松林優一朗	東京大学大学院 情報理工学系研究科	博士課程	情報抽出	全期間
小嶋大起	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	全期間
竹内淳平	東京大学大学院 情報理工学系研究科	修士課程	情報検索	全期間
岡野原大輔	東京大学大学院 情報理工学系研究科	修士課程	機械学習	全期間
深町佳一朗	東京大学大学院 情報理工学系研究科	修士課程	情報抽出	全期間
田谷滋規	東京大学大学院 情報理工学系研究科	修士課程	情報抽出	全期間
NguyenLuu Thuy Ngan	東京大学大学院 情報理工学系研究科	修士課程	機械翻訳	全期間
佐藤学	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 17 年 11 月～ 平成 18 年 3 月
中西紘子	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 17 年 11 月～ 平成 18 年 3 月
三浦研璽	東京大学大学院 情報理工学系研究科	修士課程	オントロジ学習	平成 17 年 11 月～ 平成 18 年 3 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 17 年 11 月～ 平成 18 年 8 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 17 年 11 月～ 平成 18 年 8 月
沈垣甫	東京大学大学院 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 11 月～ 平成 18 年 8 月

(2) オントロジグループ

氏　名	所　属	役　職	研究項目	参加時期
中川裕志	東京大学 情報基盤センター	教授	テキストからのオントロジ自動獲得	平成 17 年 11 月～ 平成 18 年 8 月
吉田稔	東京大学 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 17 年 11 月～ 平成 18 年 8 月
二宮崇	東京大学 情報基盤センター	講師	情報抽出	平成 17 年 11 月～ 平成 18 年 8 月
清田陽司	東京大学 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 11 月～ 平成 18 年 8 月

星野綾子	東京大学大学院 情報学府	修士課程	WEB からの情報抽出	平成 17 年 11 月～ 平成 18 年 8 月

７．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18年3月 20 日 ～21 日	Text mining, Ontologies and Natural Language Processing in Biomedicine	英国マンチェスター大学カンファレンスセンター	18 人	情報科学と生命科学という、違った分野の研究者およそ 30 名が集まり、生物医学分野における情報技術の発展について意見交換を行った

(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
鶴岡　慶雅 （英国マンチェスター大学 研究員）	構文解析器の共同開発と専門用語認識に関する研究打合せを行うため	フォレスト本郷	H18年7月8日 ～H18年7月23日

８．発展研究による主な研究成果

(1)論文発表（英文論文　5　件　　邦文論文　1　件）

- Kim, Jin-Dong and Jun'ichi Tsujii. (2006). **Corpora and their Annotation**. In Sophia Ananiadou and John McNaught, (Eds.), Text Mining for Biology and Biomedicine. 46 Gillingham Street, London SW1V 1AH UK. Artech House.
- Spasic, Irena, Sophia Ananiadou, Jun'ichi Tsujii. (2005). **MaSTerClass: a case-based reasoning system for the classification of biomedical terms**. Bioinformatics. 21(11). pp. 2749-2758. Oxford University Press.
- Ninomiya, Takashi, Yoshimasa Tsuruoka, Yusuke Miyao, Kenjiro Taura and Jun'ichi Tsujii. **Fast and Scalable HPSG Parsing.** Traitement automatique des langues (TAL). 46(2). Association pour le Traitement Automatique des Langues, 2006.
- Ananiadou, Sophia, Douglous Kell and Junichi Tsujii. **Text Mining and its potential applications in systems biology.** Trends in Biotechmology. 24(12). Elsevier, 2006.
- Yusuke Miyao and Jun'ichi Tsujii. **Feature Forest Models for Probabilistic HPSG Parsing.** Computational Linguistics. In submission.
- 岡野原　大輔, 辻井潤一 **レビューに対する評価指標の自動付与** 自然言語処理, vol.14, No.3, 2007. (to appear)

⑵口頭発表

①学会

国内　６　件,　　海外　１７　件

- Chun, Hong-woo, Yoshimasa Tsuruoka, Jin-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**. The Pacific Symposium on Biocomputing (PSB) pp. 4-15.

- Hiroshi Nakagawa and Hidetaka Masuda: **Extracting Paraphrases of Japanese Action Word of Sentence Ending Part From Web and Mobile News Articles**, AIRS2004, Lecture Note of Computer Science 3411, pp.94-105, 2005, Springer-Verlag Berlin Heidelberg

- 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, Vol.3248,pp.338-347, 2005, Springer Berlin Heidelberg

- Okanohara, Daisuke and Kunihiko Sadakane. **Practical Entropy-Compressed Rank/Select Dictionary.** In the Proceedings of ALENEX 2007. New Orleans, Lousiana, January 2007.

- Yoshida, Kazuhiro. **Ambiguous Part-of-Speech Tagging for Improving Accuracy and Domain Portability of Syntactic Parsers.** In the Proceedings of the Twentieth International Joint Conference on Artificial Intelligence. January 2007.

- Matsuzaki, Takuya, Yusuke Miyao and Jun'ichi Tsujii. **Efficient HPSG Parsing with Supertagging and CFG-filtering.** In the Proceedings of the Twentieth International Joint Conference on Artificial Intelligence. January 2007.

- Ninomiya, Takashi, Takuya Matsuzaki, Yoshimasa Tsuruoka, Yusuke Miyao and Jun'ichi Tsujii. **Extremely Lexicalized Models for Accurate and Fast HPSG Parsing.** In the Proc. of EMNLP 2006. Sydney, Australia, pp. 155--163, July 2006.

- Daisuke, Okanohara, Yusuke Miyao, Yoshimasa Tsuruoka and Jun'ichi Tsujii. **Improving the Scalability of Semi-Markov Conditional Random Fields for Named Entity Recognition.** In the Proceedings of ACL 2006. Sydney, Australia, July 2006.

- Miyao, Yusuke, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya and Jun'ichi Tsujii. **Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases.** In the Proceedings of COLING-ACL 2006. Sydney, Australia, pp. 1017--1024, July 2006.

- Sato, Manabu, Daisuke Bekki, Yusuke Miyao and Jun'ichi Tsujii. **Translating HPSG-style Outputs of a Robust Parser into Typed Dynamic Logic.** In the Proceedings of COLING-ACL 2006 Poster Session. Sydney, Australia, July 2006.

- Ohta, Tomoko, Yuka Tateisi, Jin-Dong Kim, Akane Yakushiji and Jun-ichi Tsujii. **Linguistic and Biological Annotations of Biological Interaction**

Events. In the Proceedings of the Fifth International Conference on Language Resource and Evaluation (LREC 2006). Genoa, Italy, pp. 1405--1408, May 2006.

- Yakushiji, Akane, Miyao, Yusuke, Ohta, Tomoko and Tateisi, Yuka and Tsujii, Jun'ichi. **Automatic Construction of Predicate-argument Structure Patterns for Biomedical Information Extraction.** In the Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia, pp. 284--292, July 2006.

- Ohta, Tomoko, Yusuke Miyao, Takashi Ninomiya, Yoshimasa Tsuruoka, Akane Yakushiji, Katsuya Masuda, Jumpei Takeuchi, Kazuhiro Yoshida, Tadayoshi Hara, Jin-Dong Kim, Yuka Tateisi and Jun'ichi Tsujii. **An Intelligent Search Engine and GUI-based Efficient MEDLINE Search Tool Based on Deep Syntactic Parsing.** In the Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. Sydney, Australia, pp. 17--20, July 2006.

- Unno, Yuya, Takashi Ninomiya, Yusuke Miyao and Jun'ichi Tsujii. **Trimming CFG Parse Trees for Sentence Compression Using Machine Learning Approaches.** In the Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, Australia, pp. 850--857, Association for Computational Linguistics, July 2006.

- Tateisi, Yuka, Yoshimasa Tsuruoka and Jun'ichi Tsujii. **Subdomain adaptation of a POS tagger with a small corpus.** In the Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology. New York, USA, pp. 136–137, June 2006

- Kenji Sagae, Yusuke Miyao, and Jun'ichi Tsujii. **HPSG Parsing with Shallow Dependency Constraints.** In the proceedings of the ACL 2007. Prague, Czech Republic, June 2007.(to appear)

- Okanohara, Daisuke, Jun'ichi Tsujii. **A Discriminative Language Model with Pseudo-Negative Samples.** In the proceedings of the ACL 2007. Prague, Czech Republic, June 2007.(to appear)

- 岡野原大輔 辻井潤一. **潜在的情報を利用した文識別モデル.** 言語処理学会第 13 回年次大会発表論文集. 3 月 2007.

- 松原勇介, 秋葉友良, 辻井潤一. **最小記述長原理に基づいた日本語話し言葉の単語分割.** 言語処理学会第 13 回年次大会発表論文集. 2007.

- 荒木伸夫, 吉田和弘, 鶴岡慶雅, 辻井潤一. **囲碁における正確な着手予測のためのファジーパターンマッチング.** 人工知能学会全国大会(第 20 回)論文集. June 2006

- 藤本宏涼, 國安結, 中川裕志, 吉田稔, 清田陽司: 用例検索システム Kiwi の知識テキストマイニングツールへの拡張, 言語処理学会第 12 回大会 C3-6、2006

- 海野裕也, 二宮崇 , 宮尾祐介, 辻井潤一: **機械学習を用いた文脈自由規則の書き換えによる文圧縮, 言語処理学会第 12 回大会** C5-6, 2006

- 佐藤学, 戸次大介, 宮尾祐介, 辻井潤一: **頑健な HPSG パーザの出力の TDL 意味表現への変換, 言語処理学会第 12 回大会** D5-1, 2006

②その他
国内　　件,　　海外　　件

(3)特許出願（本研究に係わり、JST から出願したものとで研究機関から出願したもの）

出願元	国内（件数）	海外（件数）
JST	0	0
研究機関	0	0
計	0	0

(4)その他特記事項

[A] 研究成果は、論文発表だけでなく、以下のサイトからソフトウェア、アノテーション・コーパス（GENIA コーパス）の形で広く公開し、国内外の研究者によって使用されている。

東京大学・辻井研究室：http://www-tsujii.is.s.u-tokyo.ac.jp/index-j.html
英国国立テキストマイニングセンター（マンチェスター大学）：http://www.nactem.ac.uk/
国立遺伝学研究所ゲノムネット公式ウェブサイト：http://www.mext-life.jp/genome

[B]本プロジェクトに関連した招待講演・チュートリアル・基調講演を多数行っている。内外において本研究プロジェクトの成果が広く認められていることを示している。

（宮尾祐介）
講演：チュートリアル
学会名： European Summer School in Logic, Language and Information(ESSLL I 2006)
開催場所：スペイン・マラガ
開催時期：2006 年 7 月 31 日～2006 年 8 月 11 日

（宮尾祐介、二宮崇）
講演：チュートリアル
学会名：言語処理学会第１３回年次大会（NLP2007）
開催場所：滋賀県、龍谷大学
開催時期：2007 年 3 月 19 日～2007 年 3 月 23 日

(金振東)
講演：チュートリアル
学会名： The 2nd International Symposium on Semantic Mining in Biomedicine(SMBM 2006)
開催場所：ドイツ・イエナ
開催時期：2006 年 4 月 9 日～2006 年 4 月 12 日

(鶴岡慶雅)
講演：チュートリアル
学会名： International Joint Conferences on Artificial Intelligence (IJCAI-07)
開催場所：インド・ハイデラバード
開催時期：2007 年 1 月 6 日～2007 年 1 月 12 日

（辻井潤一）
講演：基調講演
学会名：UK Bioinformatics Forum(UKBF)
開催場所：イギリス・ロンドン

開催時期：2005 年

講演：招待講演
学会名：International Symposium on Language and Brain by COE of the University Tohoku
開催場所：仙台
開催時期：2005 年

講演：招待講演
学会名：Symposium on Semantic Enrichment
開催場所：英国・ケンブリッジ
開催時期：2006 年

講演：パネリスト
学会名：International Committee on Computational Linguistics and the Association for Computational Linguistics(Coling/ACL-06)
開催場所：オーストラリア・シドニー
開催時期：2006 年 7 月 17 日～2006 年 7 月 21 日

講演：基調講演
学会名：The 17th European Conference on Principles and Practice of Knowledge Discovery in Databases(ECML/PKDD-06) Workshop on Data and Text Mining for Integrative Biology
開催場所：ドイツ・ベルリン
開催時期：2006 年 9 月 18 日～2006 年 9 月 22 日

講演：招待講演
学会名：The 7th International Conference on Systems Biology(ICSB)
開催場所：横浜
開催時期：2006 年 10 月 9 日～2006 年 10 月 13 日

講演：招待講演
学会名：Inauguration Ceremony of MIB
開催場所：英国・マンチェスター
開催時期：2006 年

講演：基調講演
学会名：The 20th Pacific Asia Conference on Language, Information and Computation(Paclic 20)
開催場所：中国 Wuhan
開催時期：2006 年 12 月

講演：招待講演
講演場所：Xerox Research Centre Europe、フランス・グルノーブル
講演時期：2007 年

講演：招待講演
講演場所：スイス、チューリッヒ大学
講演時期：2007 年

講演：招待講演
学会名：The BioCreAtIvE(Critical Assessment for Information Extraction in Biology)

開催場所：スペイン・マドリッド
開催時期：2007 年 4 月 23 日～2006 年 4 月 25 日

[C]研究代表者・辻井は、2005 年より英国・マンチェスター大学に設立された国立･テキストマイニングセンター(National Centre for Text Mining:NaCTeM)の所長に任命される。センターの業務は、CREST/SORST での研究と重なり、本研究での成果も、同センターを通して世界の研究者にアナウンスされている。

９．結び

［目標の達成度］

２章「研究の実施経過」の項で述べたように、３年計画で開始された本プロジェクトは、科学研究補助金・特別推進研究が、平成１８年度から開始されたことで、１年足らずで大幅な組織の改編を行い、１年５ヶ月でプロジェクトを辞退することになった。この改編に伴い、当初目標の中で、計算機環境インフラ技術、知識・意味に関する処理を汎用化する技術に関連する研究は、５年のプロジェクトである特別推進研究に引継ぎ、本発展研究は、残り７ヶ月間で達成可能と考えられる統合システムの構築、および、それに直接関連する要素技術に限定した研究を行うこととなった。

したがって、本報告書も、この変更された目標に整合的な形で作成され、数ヶ月で特別推進に移行することになった分担グループの研究成果は入れていない。

以上のような計画の変更にも関わらず、また、１年５ヶ月の短期的なプロジェクトであったにも関わらず、本研究は、豊かな研究成果をあげることができた。これは、本発展研究が、先行した５年間のＣＲＥＳＴ研究の継続研究として実施されたことから、プロジェクト構成員の問題意識とプロジェクト全体の方向がプロジェクト開始段階で十分に整合的なものであったことが大きい。

また、当初の３年計画においても、実ユーザ（生命科学者）との緊密な連携を主目標としていたこと、このために、中間時点（平成１８年度終了時点）で第一バージョンの統合システムの構築を終了し、これをプラットフォームとして連携する計画であったことから、特別推進研究発足に伴う計画の変更がそれほど大きなものにならなかった。

４章の第一節の３つのシステムは、いずれも、実ユーザからのフィードバックを得るための課題解決型のシステムであり、当初計画からのものである。遺伝研究所、マンチェスター大学バイオ研究所、産業総研の研究者からの反応もよく、今後、特別推進研究を進めていく上での有効なプラットフォームが構築できたと考えている。

４章の第２節の要素技術は、いずれも、第一節のシステムの性能を向上させる技術であり、処理速度（単純な文法を使ったものよりも高速)、処理精度ともに現時点で世界の最高水準にある。これらの要素技術は、特別推進研究の中でさらに発展させることで、言語処理の次世代基盤技術を確立できると思っている。

４節の第３節のコーパス、特に、事象アノテーションつきのコーパスは、世界で最初のものであり、意味・知識に基づく処理の基盤を構築したものである。この成果自体は、現時点では、まだ、第一節・第二節のシステム、要素技術とは結びついていない。ただ、実ユーザからフィードバックの中で、一番多いものが抽象的な生命事象に基づく検索を可能にするという要望であり、事象アノテーションつきの GENIA コーパスは、次世代 MEDIE（４．１．１節参照）の構築に大きな貢献をなすと思っている。これも、特別推進研究に引き継がれる。

以上のように、本プロジェクトの成果の多くは、特別推進研究に引き継がれるが、３つの課題解決型システムの構築、英語解析プログラムの高効率化と分野適応技術など、本研究の成果としてまとまりのある成果を上げることができたと考えている。

［成果の意義と自己評価］

本研究に先行する CREST を開始した当初は、テキストマイニング、情報抽出、知的テ

キスト検索といった実世界の応用システムに、HPSG のような言語理論に基づく言語処理技術が使われると考える研究者は、皆無であった。それが、CREST/SORST という６．５年間の研究の結果、巨大なテキスト集合を対象とした検索システムに有効に活用できること、また、病疾患と遺伝子の関係マイニングや蛋白質相互関係の情報抽出の性能を向上させることが実際に示せたことは、非常に大きな意義があったと思う。

生命科学のテキストマインニング分野、あるいは、情報抽出の分野では、ここ１年の間に文解析を使った研究が発表されるようになり、我々の主張が正しかったことが認められつつある。ただ、これらの論文も、CFG という比較的単純な文法枠組みを使うにとどまっており、我々との技術の差は大きい。我々が、世界との技術レベルでこれだけの差をつけることができたのは、5 年間という長い研究期間を与えてくれた CREST プロジェクトと、研究グループの継続性を保つことを可能にした SORST というプロジェクト枠組みがあったためだと思っている。プロジェクトが短期指向、応用指向になっている米国・ヨーロッパに比較して、恵まれていたと考えている。

［プロジェクト運営］

情報技術の研究は、優秀な研究者の集団を構築できるか、どうかが成否の鍵となる。CREST 研究に参加した研究者の何人かが、CREST 終了時点で、大学（東京大学、マンチェスター大学、工学院大学など）での研究者として雇用された。　ただ、彼らも、SORST 研究推進に引き続き協力し、かつ、新たに優秀な研究者が雇用できたことが、本研究を成功させる大きな要因となった。研究に参加していた（あるいは、いる）若手研究者は、ESSLLI・IJCAI・SMBM など、国際的な会議で招待講演、チュートリアル講演を行うなど、世界をリードする研究者として育っている。

また、情報技術研究では、実ユーザと密接な連携をとり、彼らの要求を研究課題としてうまく定式化していくことも、研究成功の重要な鍵となる。この点では、国立遺伝学研究者・五條堀教授、産業総研・今西博士、マンチェスター大学・D.Kell 教授、理化学研究所・林崎博士など、実ユーザのグループと緊密な連携関係が確立できたことは、プロジェクトの推進上、大きな助けとなった。

［今後の研究方向］

テキスト処理における量と質の技術の統合は、テキスト処理を知識発見と知識管理の技術という、情報技術の中核技術にテキスト処理を融合することになる。知識の中に占めるテキストの重要性を考えると、大量テキストと知識とを結びつけるテキストマイニング技術は必須の技術となる。

本研究では、生命科学、とくに、分子生物学でのＴM技術に焦点を当てたが、開発した技術の枠組みは一般的なものであり、広範な分野でのＴMに適用可能である。また、本研究では、言語処理技術を中心とした研究を行ったが、これを数値データ・画像データなどからのマインニング技術と組み合わせることで、将来、さらに大きな研究分野へと発展するものと考えている。