「高度メディア社会の生活情報技術」
平成12採択研究代表者

辻井　潤一

（東京大学大学院情報学環　教授）

「情報のモビリティを高めるための基盤技術」

1.研究実施の概要

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

過去４年間で、(1) 意味表現を出力する文構造解析プログラム（Enju）、(2)テキスト構造や意味に基づく知的な索引構造、(3)検索過程での推論の起動とオントロジー表現、(4)高効率で高耐性なクローラ、(5)オントロジーの半自動的構築とテキストからの情報抽出手法、(6)生命科学のテキスト情報処理のためのリソース(GENIAコーパス)の構築など、中核となる要素技術・プログラムを完成させた。

最終年度では、これらの研究成果を統合することで、我々が開発してきた要素技術の有効性を実証する。

2.研究実施内容

[目的]

大量テキストからユーザが必要とする情報を含むテキストの部分を同定する質問応答(Q／A)システムの研究は、米国を中心に活発化している。また、ユーザの質問の答えを含んだ複数のテキストから冗長な部分を取り除き、整合性のある一つのテキストとして提示する野心的な技術も、情報融合(Information Fusion)の技術として研究が開始されている。

しかし、米国・DARPAが進めるこれらの研究は、システム構築を急ぐあまり、アドホックな技術の集積となり、見通しのよい、しかも、異なる主題分野へと移行可能なGenericな基盤技術の研究とはなっていない。たとえば、研究用プロトタイプとして作成されているQ/Aシステムの大部分は、キーワードによる検索をGoogleなど既存の検索エンジンを用いて、まず対象テキストを限定し、第２段階で言語処理技術を用いた処理を行う。このため、対象テキストを限定する第一段階には、言語処理、知識処理の成果は全く使えない。また、第２段階の言語処理も、実時間処理の時間的な制約から本格的な技術は使用できず、

パターン照合など非常に単純な処理が使われている。知識に基づく推論処理、深い意味に基づく言語処理の重要性は強調されているが、研究の実態は単純な技術の適用にとどまっている。

これらの欠陥は、

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

という、研究枠組み自体の不備によるものである。

本プロジェクトでは、言語処理、分散処理ソフトウェア、エージェント技術、オントロジー研究という、異なった分野の研究グループが連携をとることで、テキストの収集、検索、知識処理、テキストからの情報抽出など、関連技術を有機的に統合したシステムを開発する。このような統合的な研究は世界でも全く行われておらず、９０年代後半に研究された情報抽出の技術など、Genericな言語処理、知識処理の技術を、知的なQ/Aの基盤技術として確立することを目指す。

[方法]

プロジェクト前半の３年間で、以下の７つの要素的基盤技術の分野で革新的な技術を開発し、後半の２年間でそれらを有機的に統合する。

(1)基盤システムの構築

言語処理、知識処理の結果をText Annotationとして捕らえ、これをテキスト集合への索引構造に反映することで、テキスト構造、文の意味を基にしたテキスト検索を高速に行うことができるシステムを構築する。これは、言語処理、知識処理の成果をテキスト検索に統合するための基盤システムとなる。

このシステムの基本として、

(a)領域代数(Region Algebra)に基づくタグ付きテキストの領域検索のモデル

(b)タグを素性構造表現に拡張し、タグ間の整合性を単一化操作とする検索モデル

を採用し、これらのモデル上の検索処理を高速に実行するための索引構造を設計、開発する。

(2)情報抽出のための言語処理技術の開発

情報抽出の技術は、パターン照合や有限状態オートマトン(FSA)など、単純な言語処理モデルが使われてきた。しかしながら、意味や知識を考慮した、より精度の高い処理を取り込むためには、言語学的に正当な枠組みを用いて、表層の統語構造と論理的な意味構造とを系統的に結びつけることが必要となる。本プロジェクトでは、

(a)実テキストからの文法規則の自動獲得に関する研究

(b)実テキストからの確率モデルの学習に関する研究

という２つの研究を行い、文法の被覆率（Coverage）を向上させると同時に、曖昧さ(Ambiguity）の増大に対処し、出力する意味構造の数を減少させ、知的検索という実用場面で使える高効率で、高耐性な言語処理技術を完成させる。

(3)特定分野のオントロジー、Annotated Textsの構築

オントロジーに基づく推論処理と知的な情報検索に結びつけるために、特定分野の詳細な知識を実際に構築する。また、大量のテキストを人手で分析し、分野の中核をなす意味カテゴリーとその相互関係を認定する。テキストの人手による分析結果は、Annotated Textとして計算機管理する。このAnnotated Textを用いた知識の自動構築、専門用語(Terms）の自動認識技術を開発する。

特に、本研究プロジェクトでは、実ユーザが存在し、かつ、大規模な知識ベースの基盤があるゲノム生命科学を取り上げ、分野専門家と緊密に協力することで、テキスト分析と分野オントロジーの作成を行う。

(4)高効率で高耐性なCrawlerの開発

Webクローラが行っている様々な処理のうち，実際にウェブページのダウンロードを行う部分を，再利用の容易なコンポーネントとして独立に実装し，その高速化，ならびに単独での性能評価を行う．ひとつのウェブサーバへのアクセスを集中的に行わない，Robot Exclusion Protocolの遵守を行うことで，大量のWebデータの収集を必要とする研究者が，簡便に利用できるツールを構築する．そして，それを簡単なスクリプト処理によってコントロールし，用途によって収集の戦略を様々に変えながらページを収集できるインタフェースを設計する．

（２）に関して，分散管理された多数のクラスタを同時(並列)に，容易に利用できるツールを開発する．それを用いて，生物学テキストやクローラが収集した大量のデータを，多数(300CPU以上)の計算機で並列に処理する道具を提供する． また，このようなデータ集約的な計算において，データの位置などをユーザが意識しなくてもすむ，簡単化された，データマイニング・テキストマイニングの支援システムを設計する．簡単な文法によって，既存ツールを組みあわせて，分散した大量のデータを処理することができる．

(5)ユーザの空間・身体的知覚を考慮した情報の提示方式の研究

ユーザの空間的・身体的な知覚を考慮に入れた情報提示方式について研究する。このため、個人の文脈に適合した意味構造構築機能をもつ知識アーカイブ管理支援システムの開発、没入空間における知識インタラクションのためのジェスチャインタフェースの開発、非言語情報に重点をおいた実世界エージェントとのインタラクションの方式の開発を行う。

(6)Webページの表形式データからの情報抽出

表形式から抽出されたオントロジーを用い、WWWから、オントロジーに記述された分野に適合する文書を収集するための手法を開発する。WWWという統制されない文書集合から、対象分野の文書のみを集めることにより、オントロジーを用いた知的処理に適したコーパスを自動生成することを目指す。

表形式データの構造を統計的手法で認識し、属性と属性値のペアの形で抽出する方法の

確立を目指す。

(7)WWWからの言語横断的な情報抽出

WWWにおける多数の言語で記述されたWebページから用例検索、知識マイニングを行う方法の確立を目指す。また、この方法をWWW上に公開されないローカルなファイルに適用して、そのファイルの特有の知識をマイニングできるツールの構築を目指す。

多言語にわたる情報抽出のうち日本語に特有の問題である西欧言語起源のカタカナ語の表記の揺れを認識、同定するシステムを目指す。

[研究の経過と成果]

要素技術として設定した７つの基盤技術に関して、過去４年間の研究で非常に大きな効果を挙げてきた。その主なものを次に列挙する。

(1)HPSGによる英文解析システム：

言語学的に妥当な文法に確率モデルを定義することに成功し、２億語からなる英文テキストの解析を行った。これは、表層文から意味構造を出力するシステムで、このような大規模データを処理できるものとしては世界最初のものである。文法の精緻化を行うことにより、その精度は、現在、世界でも有数のものとなっている（表１）。

(2)高効率な領域代数処理システム：

埋め込み型のタグをもつXMLデータに対しても高効率で検索を行う領域代数システムを開発。（１）で作成された２億語のテキストの意味構造から、指定の意味構造の検索が１秒以下で可能なことを確認。

(3)生命科学のAnnotatedText（GENIAコーパス）の作成：

２０００抄録（２万文、５０万語）のAnnotatedTextを作成し、世界に公開（図１）した。現在、２４０を越える研究チームがこのデータを使って研究。また、このデータは、いくつかの国際ワークショップでのゴールド・スタンダード・データとして活用される。GENIAコーパスは、意味タグのついた世界最大規模のコーパスであるだけではなく、文の構文構造・名詞句間の参照関係・生命現象にかかわる事象など、非常に豊かな情報が付与されている。

(4)英文解析システムの分野依存的な学習手法：

一般分野で学習した英文解析システムの確率モデルを特定分野のコーパスに適用することにより、分野に依存した文構造を有効に解析できることを実証した。これは、分野とタスクに依存した言語処理システムを構築するための基盤技術となる。

(5)分野に依存したPOSタガー、NERの開発：

MEMMの新しい最適化手法として、双方向・最尤探索の手法を提案し、それがCRF(Conditional Random Field)と同様に大域的な最適化の効果をもち、かつ、CRFに比べてはるかに高速に実装できること、また、それが英語品詞付与やNERに有効であることを実証した。さらに、この手法を英語のPOSタガーや生命科学のNERに適用して、従来手法よりも優れた精度がでることを確認した。

(6)分散処理ソフトウェアの言語処理への活用：

ソフトウェア研究グループが開発した技術を使って、文解析システムの実験を行い、２年分のMedline抄録（２億語）を３０時間で処理（表２）。PCクラスタによる分散処理を行わない場合には、２週間の処理時間。現在、テキスト規模を１０倍にした場合の実験を計画。

(7)Crawlerによるテキスト収集とその活用：

平均40ページ/秒、短い期間の実行においては、200ページ/秒程度のクローリング速度を達成。2ヶ月間の稼動で実際に2億ページ弱を収集。この収集されたテキストを活用した機械学習実験、テキスト検索実験を行った。

(8)Crawlerの大幅な高速化手法：

ひとつのウェブサーバへのアクセスを集中的に行うことなく，1台の計算機で600ページ/秒のダウンロード速度を数時間維持することに成功した．また，そのcrawlerを20台程度で並列に実行し，この台数まで性能がスケーラブルに向上する(約10,000ページ/秒)ことを確認した．

(9) PCクラスタを有効に活用するためにツール（GXP）の開発：

GX`Pを実応用に適用し、10以上のクラスタ，1300台以上の計算機で快適に動作させることができた．自然言語処理グループによって利用され，そこでは,crawlerによって収集した約4億のwebページから日本語文(5億文)抽出，CFG構文解析などを行う処理を，4つのクラスタ/SMP計算機(合計約350CPU)を用いて並列処理を行った．これにより，1000 CPU日程度の量の処理を4，5日程度で完了することができた．

以上のように、基盤となる要素技術の開発は非常に順調に進展した。現在、後半の２年間でこれらの成果を統合したシステムを開発するため、実ユーザである生命科学の研究グループと緊密な共同研究を行っている。

	LP	LR	UP	UR
ベースライン	*77.57*	*76.43*	*82.20*	*80.99*
Enju	*88.26*	*87.11*	*91.51*	*90.32*

LP, LR, UP, URは、それぞれ、Labeled Precision, Labeled Recall, Unlabeled Precision, Unlabeled Recallを意味する。
実験には、Wall Street Journalを用いた。

［表１］ 現在の解析精度

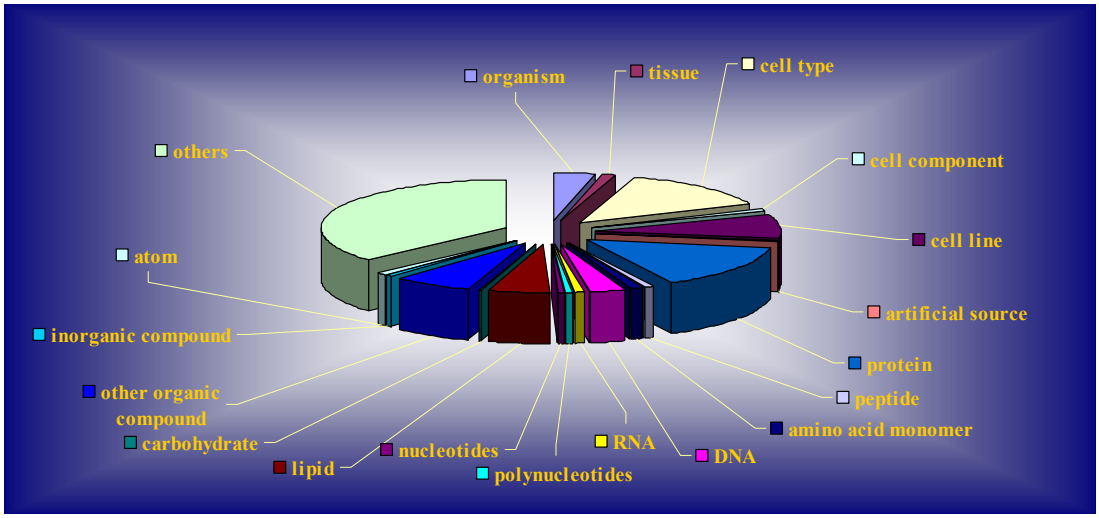

図１　GENIAコーパス中での意味クラスの分布

［実行環境］

PCクラスタ　（合計350プロセッサ）

IBM BladeCenter Cluster

Dual Xeon 2.4GHz × 69　＋　Dual Xeon 2.8GHz × 42

Appro Blade Server

Dual Xeon 2.4GHz × 64

並列実行プラットフォーム： GXP (Grid Explorer)

［実験データ］

MEDLINE 118万論文　(2000年1月～2002年3月)

規模： *9百万文（2億2千万語）*

処理時間：３０時間

表２　並列分散処理による文解析実験の結果

3.研究実施体制

(1) 言語処理グループ

① 研究分担グループ長：辻井潤一（東京大学大学院情報学環、教授）

② 研究項目：情報のモビリティを高めるための言語処理技術の研究および全体の統括

(2) 広域・分散ソフトウエアグループ

① 研究分担グループ長：米澤明憲（東京大学情報理工学系研究科、教授）

② 研究項目：言語処理のための大規模データ収集およびサービスシステム

(3) オントロジーグループ

① 研究分担グループ長：中川裕史（東京大学情報基盤センター、教授）

② 研究項目：オントロジーのための記述枠組とオントロジー構築に関する研究

(4) エージェント対話グループエージェント対話グループ

① 研究分担グループ長：西田豊明（京都大学情報学研究科、教授）

② 研究項目：ユーザの空間・身体的知覚を考慮した情報の提示方式の研究

4.主な研究成果の発表

(1) 論文発表

○ Tsujii, Jun'ichi. (2004)

Thesaurus or logical ontology, which do we need for mining text? (keynote speech)

In the Proc. of Language resources and evaluation conference (LREC 2004). Vol.III. 55-57　rue Brillat Savarin 75013 Paris France. pp. pp IX-XVI. ELRA.

○ Miyao, Yusuke and Jun'ichi Tsujii. (2004)

Deep Linguistic Analysis for the Accurate Identification of Predicate-Argument Relations

In the Proceedings of COLING 2004. pp. 1392-1397.

○ Tateisi, Yuka and Jun'ichi Tsujii. (2004)

Part-of-Speech Annotation of Biology Research Abstracts

In the Proceedings of 4th International Conference on Language Resource and Evaluation (LREC2004). IV. pp. 1267-1270.

○ Yakushiji Akane, Yuka Tateisi, Yusuke Miyao and Jun'ichi Tsujii. (2004)

Finding Anchor Verbs for Biomedical IE Using Predicate-Argument Structures

In the Companion Volume to the Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics. pp.157-160

○ Tsuruoka, Yoshimasa and Jun'ichi Tsujii. (2004)

Improving the Performance of Dictionary-based Approaches in Protein Name

Recognition.
In the Journal of Biomedical Informatics. 37(6). pp. 461-470. Elsevier.

○ Chun Hong-Woo, Tomoko Ohta, Jin-Dong Kim and Jun'ichi Tsujii. (2004)
Building Patterns for Biomedical Event Extraction
In the 15th International conference on Genome Informatics (GIW)-Poster and Software Demonstrations. 15(2). pp. 163. Universal Academy Press, INC.

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

○ Takashi Tajima, Yong Xu, Toyoaki Nishida(2004)
Entrainment Based Human-Agent Communication of Tacit Intention
In the Proceedings of International Workshop on Intelligent Media Technology for Communicative Intelligence (IMTCI 2004), Warsaw, September, 2004.

○ Hung-Hsuan Huang, Yasuyuki Sumi, and Toyoaki Nishida(2004)
Gallery: In support of human memory
In M.Gh.Negoita, R.J.Howlett, L.C.Jain (Eds), 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2004), LNAI 3213, pp.357-363, September, 2004.

○ Burin Anuchitkittikul, Masashi Okamoto, Sadao Kurohashi, Toyoaki Nishida,Yoichi Sato(2004)
Video Content Manipulation by Means of Content Annotation and Nonsymbolic Gestural Interfaces
In M.Gh.Negoita, R.J.Howlett, L.C.Jain (Eds), 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2004), LNAI 3213, pp.385-392, September, 2004.

○ Kenji Kaneda, Kenjiro Taura, Akinori Yonezawa (2004)
Routing and Resource Discovery in Phoenix Grid-Enabled Message Passing Library
In 4th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid'04).

○ 安藤雅享, 田浦健次朗, 近山隆(2004)
Grid環境での並列ジョブ投入を支援するシェル
SACSIS 2004 - 先進的計算基盤システムシンポジウム

○ 山田 雅信, 田浦 健次朗, 近山 隆, 高橋 俊行 (2004)
インクリメンタルPageRankによる重要Webページの効率的な収集戦略

情報処理学会論文誌(トランザクション)コンピューティングシステム 2004年10月
Vol.45