「高度メディア社会の生活情報技術」
平成12年度採択研究代表者

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３年間で、(1) 意味表現を出力する文構造解析プログラム、(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ー、(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など、中核となる要素技術・プログラムを完成させた。今後２年間で、これらを統合することでその有効性を確証する。

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５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

PCクラスターを用いた大規模データの収集システムの設計、性能向上を行い、実際にウェブデータの収集と蓄積を行う。クローリングを行う計算機資源として、IA32のLinuxをノードとする、多数ノードからなるクラスタシステムを用いる。また、収集された大量データを利用した情報抽出や自然言語処理の実験を行うため、データサーバが備えるべき抽象化、APIや、実行方式を設計、開発する。

収集目的を収集過程に反映するために収集対象や順序を制御すること、集められたデータに対する処理を、分散透明かつ並列に行うことを目指した汎用のシステムを開発する。

(5) ユーザのモデルを考慮した情報の提示方式の研究

収集されたテキスト情報をユーザに提示する手法として、Visualizationなどさまざまな手法がすでに提唱されている。本プロジェクトでは、特に言語的な情報の提示に焦点をあて、発話を構成する文章を、状況や受け手の背景知識や興味に合わせて動的に生成する方式、国語辞典を用いて与えられた文章を平易な文章に通常の書き言葉様式のテキストから、会話型の質問・応答形式の非線状的な系列を生成する方式、大量の情報の中から重要な意味関係を抽出し、それを簡潔な自然言語に言い換える方式を中心とした研究開発を行う。

[研究の経過と成果]

要素技術として設定した５つの基盤技術に関して、過去３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 （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)。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。

(4)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。

(5)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。

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	*76.9*	*76.8*	*80.7*	*80.6*
既存のモデル	*80.6*	*81.0*	*84.8*	*85.1*
Feature forest	*86.9*	*86.9*	*89.6*	*89.5*

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 Sreet Journalを用いた。

［表１］　現在の解析精度

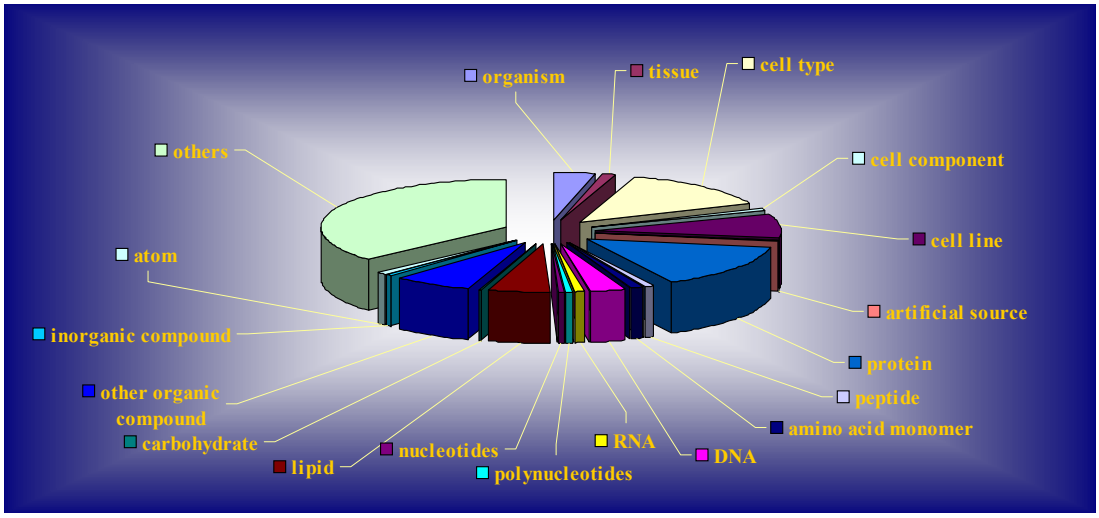

図１　GENIAコーパス中での意味クラスの分布

―――――――――――――――――

［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

① 研究分担グループ長：西田豊明

② 研究項目：情報提示とインターラクションに関する研究

4.主な研究成果の発表（論文発表および特許出願）

（１）論文（原著論文）発表

○ Miyao, Yusuke, Takashi Ninomiya and Jun'ichi Tsujii.

Lexicalized Grammar Acquisition.

In the Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL) companion volume. 127―130,

2003

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

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

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

○ 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. pp. 58--60.

○ Yakushiji, Akane, Yuka Tateisi, Yusuke Miyao, Naoki Yoshinaga and Jun'ichi Tsujii.
A Debug Tool for Practical Grammar Development.
In the Proceedings of the 41st ACL companion volume. pp. 173--176. 2003

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

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

○ Masuda, Katsuya.
A Ranking Model of Proximal and Structural Text Retrieval Based on Region Algebra.
In the Proceedings of the ACL 2003 Student Research Workshop. pp. 50--57.2003

○ Tsuruoka, Yoshimasa and Jun'ichi Tsujii.
Boosting Precision and Recall of Dictionary-Based Protein Name

Recognition.
In the Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine. pp. 41-48.

○ Kumiko Tanaka-Ishii, Masato Yamamoto, Hiroshi Nakagawa.
Kiwi: A Multilingual Usage Consultation Tool based on Internet Searching,
Proceedings of the Interactive Posters/Demonstrations, ACL-03, pp.105-108, Sapporo, July 2003

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

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

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

○ Kenji, Miura, Yoshimasa Tsuruoka and Jun'ichi Tsujii.
Automatic acquisition of concept relations from web documents with sense clustering.
In the IJCNLP 2004 Interactive Poster/Demo

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

○ Tsuruoka, Yoshimasa and Jun'ichi Tsujii.
Iterative CKY Parsing for Probabilistic Context-Free Grammars.
In the Proceedings of IJCNLP 2004.

○ Ninomiya, Takashi, Yusuke Miyao and Jun'ichi Tsujii.
A Persistent Feature-Object Database for Intelligent Text Archive Systems.
In the Proceedings of IJCNLP 2004.

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

○ 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 the Proceedings of IJCNLP-04.

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

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

○ Hidetaka Masuda, Shuuich Tsukamoto and Hiroshi Nakagawa
Recognition of HTML Table Structure
Proceedings of IJCNLP (International Joint Conference of Natural Language Processing) 2004, pp.183-188, Hainan-Island, China, March 2004

○ Kumiko Tanaka-Ishii, Michiko Abe, Hiroshi Nakagawa.
Categorization of movies using comments,
Proceedings of PACLING'03 (Pacific Association for Computational LINGuistics), pp.221-229, Halifax, Nova Scotia, Canada, Aug. 2003

○ Takashi Hoshino, Kenjiro Taura and Takashi Chikayama.
An Adaptive File Distribution Algorithm for Wide Area Network.
Workshop on Adaptive Grid Middleware. September 2003

○ 山田雅信, 田浦健次朗, 近山隆, 高橋俊行
インクリメンタルPageRankによる重要Webページの効率的な収集戦略
先進的計算基盤システムシンポジウム. May 2003.

○ 安藤雅享, 田浦健次朗, 近山隆.
Grid環境での並列ジョブ投入を支援するシェル
先進的計算基盤システムシンポジウム (SACSIS2004), May 2003.

（２）特許出願

H15年度特許出願件数：0件