「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理工学系研究科　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

本プロジェクトの目的は、Web などネットワーク化された情報システム中に散在する情報、とくに、テキスト情報をあらかじめ構造化しておくことにより、ユーザが必要とする情報を有効に、かつ、わかりやすい形で提示する基盤技術を確立することである。このために、(1)テキストという非定形な情報表現を整理された構造表現へと変換する自然言語処理技術、(2)テキスト表現と背景知識を結びつけるオントロジー技術、(3)情報ネットワークから膨大な情報を選択的に収集するソフトウェア技術、(4)ユーザの情報要求と情報提供者の意図を考慮しわかりやすい表現と形態で情報を提示するエージェント技術、の各研究を行うとともに、その統合化のための研究を行う。

プロジェクトの初期 2 年間は、研究目的の実現に向けての要素技術の研究に焦点をあて、中期 2 年間に要素技術の統合化のための研究、最終年度では総合システムの構築、研究成果の取りまとめを行う。この全体計画のなかで、平成 13 年度は、次のような要素技術の研究、および、次年度以降の大規模な技術開発を行うための資源整備を行った。

(a)　Web に散在するテキスト・非テキスト情報からの分野オントロジーの自動構築
(b)　XML/HTML による半構造化されたテキストからの情報抽出と検索
(c)　素性構造オブジェクトの高速検索手法
(d)　特定分野（ゲノム科学）の分野オントロジーと言語資源の開発
(e)　分野に適合的する学習機能をもった構文解析技術
(f)　ネットワークからのスケーラブルな情報収集方式の確立
(g)　ネットワークからの情報収集実験を行うための基盤システムの構築
(h)　背景知識を考慮したパラフレーズによる情報提示技術
(i)　言語と非言語情報の統合に関する技術
(j)　背景知識の差を解消するためのオントロジー変換技術

これらの要素技術の研究での優れた成果は、多くの国際学会で発表された。

また、ゲノム科学の研究グループとの協同でゲノム科学におけるテキスト処理と知識獲得、オントロジー構築に関する国際ワークショップを開催し、大きな成果を得た。特に、この分野が我々の成果の有効性を確証する格好のテストケースになることを確認し、我々の開発してきた各種の資源を

利用するための国際的なコンソーシアムを組織した。

2. 研究実施内容

情報システム中に散在する情報を構造化・知識化し、人間にわかりやすく提供する技術は、分野の異なる多様な要素技術の統合によってはじめて可能となる。個別な要素技術を統合したシステムを構築する以前に、Robust で斬新な要素技術を確立する必要がある。本プロジェクトの4つの技術グループは、それぞれ以下の研究を行った。

[テキスト処理、半構造化データの処理と検索](自然言語グループ)

(A) Web に散在する情報からの分野オントロジーの自動構築: Web ページを構成するためのタグ情報は、視覚的な情報提示という役割だけでなく、情報内容の論理的な構造を反映したものになっている。この要素技術では、タグによる構造化、特に、表形式の情報提示から特定分野のオントロジーを自動構築し、さらにそれをシーズとして、テキスト部分の情報を構造化する技術を開発した。次年度、大規模な実験により手法の有効性を確認する。

(B) XML/HTML による半構造化されたテキストからの情報抽出と検索:タグ情報をテキスト中の一次元領域を特性化するものとみなし、領域間の操作を定義、その操作を組み合わせることで、テキスト構造を参照した検索を可能するシステム(TIMS)を開発。この領域間の操作の定式化と Web ページの検索へと拡張する研究を行った。定式化の基礎として、カナダ・ウォータルー大学の GCL を用い、また、膨大な Web ページの検索には、発見的手法による探索空間の限定を行う手法を提案し、その有効性を確認する実験を行った。

(C) 素性構造オブジェクトの高速検索手法:型付素性構造で表現されたテキスト処理の結果を有効に検索や情報抽出に活用するために、素性構造オブジェクト中のパスに注目したインデックス構造を設計し、その効率を確認した。これは、テキストからの情報抽出の結果の蓄積、テキスト・コンテンツの知的検索、例主導(Example-based)の言語処理における例と入力との照合など、今後、統合システムの中核技術に展開させる予定である。

(D) 特定分野(ゲノム科学)の分野オントロジーと言語資源の開発:テキスト中の情報の構造化・知識化の研究には、その応用分野として、明確な情報要求を持ったユーザが存在する分野を選定する必要がある。我々は、我々の研究の有効性を実証する分野の一つとして、ゲノム科学を選定し、この分野の研究を推進するのに必要な資源の整備を行った。具体的には、我々が開発したオントロジー(GENIA オントロジー)に基づいて4000抄録(約 100 万語)のコーパスに意味のアノテーションを付与し、同時に、品詞情報など、今後の研究に必要な情報を付与した。また、この資源を有効に活用するための世界的なコンソーシアムの構築に着手した。

(E) 分野に適合する学習機能をもった構文解析技術:背景知識、分野オントロジーをテキストから構築するためには、分野の特殊性に適合的な文法、辞書を(半)自動的に構築する必要がある。このための技術として、HMM、SVM などの機械学習を使った専門用語認識システムを開発し、その有効性を(D)の資源を使って実証した。また、我々が開発してきた XHPSG の文法を使って、テキストから Sub-categorization などの語の統語的特徴を獲得する技術、統語構造の統計的な

偏りを活用する構文解析技術の基礎を確立した。

[情報収集のための高効率・高信頼ソフトウェア](広域・分散ソフトウェアグループ)

(F) ネットワークからのスケーラブルな情報収集方式の確立:並列 Crawler のアーキテクチャとして従来用いられてきたサーバ・クライアント方式が持つ欠陥を克服した、P2P 型のアーキテクチャに基づく並列 Crawler 方式の設計を行った。この並列 Crawler 方式は、既知の URL を全ノードで分散して管理することにより、従来方式のボトルネックを解消する。この方式を前提としたソフトウェア・コンポーネントを試作し、各コンポーネントの性能評価を行い、次年度の大規模並列 Crawler 開発の準備を完了した。

(G) ネットワークからの情報収集実験を行うための基盤システムの構築:Crawler の性能評価、実際のデータ収集を行うための設備として、SunBlade16 ノード、NEC Express からなるクラスターを購入、次年度以降の実験環境を整備した。

[エージェント技術に基づく有効な情報提示](エージェント・対話技術グループ)

(H) 背景知識を考慮したパラフレーズによる情報提示技術:情報提供者、情報消費者の背景知識の差を考慮し、理解しやすい表現へとパラフレーズするための基礎技術を開発した。また、国語辞典中の記述をそのままパラフレーズ規則として活用する方式を構築し、言語表現によって起動する連想・推論方式のための基本データとして、国語辞典の定義文中から述語の格枠組みを抽出する技術を開発した。

(I) 言語と非言語情報の統合に関する技術:実世界の状況に応じた有効な情報流通を行うためには、使用者の置かれた状況など、非言語的な情報と言語情報との相互関係の認識が不可欠である。このための基礎研究として、動作の特徴量から言語表現への写像を学習する HMM を構成し、これが副詞的概念を学習することを示した。

[オントロジー変換、オントロジーと情報検索](オントロジー・グループ)

(J) オントロジーアライメント技術:Yahoo、Infoseek のディレクトリ構造をオントロジーとみなし、この2つの体系間の写像を自動学習する実験を行った。具体的には、各ディレクトリ・クラスに対応する語が生起するテキスト集合でそのクラスを表現し、その類似度を計算することで最適のオントロジー変換を求めた。人間の主観的な正解集合に対して、約60%の適合度を達成することが確認された。

(K) オントロジーを利用した文書検索:検索要求と文書双方を領域オントロジーに対する写像(オントロジー・ベクトル)に変換し、このベクトル間の類似度を計算することで適合文書を検索する方式を開発した。この手法が単純なキーワードマッチングよりも、再現率・適合率ともに優れていることを実験により確証した。

3. 研究実施体制

自然言語グループ

(1) 研究分担グループ長:辻井潤一(東京大学大学院情報理工学系研究科、教授)

(2) 研究項目:テキスト処理と半構造化データの検索

(i) 高度で Robust な文構造解析手法

(ii) 情報抽出とテキスト情報の構造化

(iii) 質問応答システムと知的検索システム

広域分散ソフトウェアグループ

(1) 研究分担グループ長:米澤明憲(東京大学大学院情報学環、教授)

(2) 研究項目:Web からの情報収集のための高効率・高信頼度のソフトウェア

(i) 広域・分散型ソフトウェアの構築

(ii) 知的情報収集のための移動エージェント

エージェント・対話技術グループ

(1) 研究分担グループ長:西田豊明(東京大学大学院情報理工学系研究科、教授)

(2) 研究項目:人間にやさしい情報提示の方式

(i) テキストのパラフレーズ処理

(ii) 使用者(人間)の置かれた状況に適応的な情報提示

オントロジー・グループ

(1) 研究分担グループ長:中田圭一(東京大学大学院新領域創成科学研究科、教授)

(2) 研究項目:オントロジー工学の情報構造化への適用

(i) オントロジーの自動構築、複数オントロジーの相互変換

(ii) オントロジーの情報検索への応用

4. 発表論文

(1) 論文発表


○ N. Yoshinaga, Y. Miyao, K. Torisawa, and J. Tsujii: Resource sharing among HPSG and LTAG communities by a method of grammar conversion from FB-LTAG to HPSG, in Proc. of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education, pp39-46, Toulouse, July, 2001.

○ N.Yoshinaga, Y. Miyao : Grammar conversion from LTAG to HPSG, in Proc. of the sixth ESSLLI, pp309—324, Helsinki, Aug., 2001.

○ T. Ogure, K. Nakata, and K. Furuta : Ontology Processing for Technical Information Retrieval, in Proc. 1st International Conference on Universal Access in Human-Computer Interaction (UAHCI), New Orleans, Aug., 2001.

○ M. Yoshida, K. Torisawa and J. Tsujii: A method to integrate tables of the World Wide Web, in Proc. of the International Workshop onWDA 2001, pp31-34, Sept., 2001.

○ H. Mima, S. Ananiadou, G. Nenadic and J. Tsujii: TIMS - A Workbench for Ontology-based Knowledge Acquisition and Integration, in Proc. of Natural Languge Processing in Biomedical Applications (NLPBA 2002), Nicotia, Feb., 2002

○ Y. Miyao and J. Tsujii: Maximum Entropy Estimation for Feature Forests, in Proc. of

HLT2002, San Diego, March, 2002.

○ T. Ohta, Y.Tateisi, Jin-Dong Kim, H.Mima, J.Tsujii: GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain, in Proc. of HLT2002, San Diego, March, 2002.

○ H. Mima, S. Ananiadou, G. Nenadic, and J. Tsujii: XML Tag Information Management System ・A Workbench for Ontology-based Knowledge Acquisition and Integration, in of Proc. of Human Language Technology Conference (HLT 2002), San Diego, March, 2002

(2) 特許出願

なし