JST-9

情報のモビリティを高めるための基盤技術
Key Technologies Supporting the Mobility of Information

辻井 潤一[†‡]
Jun'ichi Tsujii

1. 背景

印刷媒体から電子的なネットワークへの情報流通形態の変化は、背景の異なる多様な主体（個人、組織）が情報を直接交換することを可能にし、物理的な情報のモビリティを飛躍的に向上させた。しかし、物理的なモビリティの向上は、必ずしも、有効な情報のモビリティが達成されていることを意味しない。情報発信者と要求者が均質な背景知識を共有しない場合には、キーワードという限定された表現から検索意図の復元を行う現在の情報検索技術は、うまく機能しない。情報要求者の個別性を積極的に取り入れた、すなわち、要求者固有のオントロジーを反映した動的な情報検索のモデルが必要となる。専門家と非専門家といったように、背景知識も言葉の使い方も全く異なる主体間の情報交換を支援するには、発信者と要求者の背景知識を調整しながら情報検索を行い、同時に、検索結果の提示も要求者の背景と文脈に依存して行うことが不可欠となる。ここでも、情報の表現自体（テキストやキーワード）とその表現を背後で支えるオントロジーの関係を考慮した処理が必要となる。

2. プロジェクト概要

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。

3. 研究成果

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ている。

1. 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 (Yoshida 2001)
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。今後、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 (Mima 2002)
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 (Ninomiya 2002)
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 (Ohta 2002)
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。

(e) 分野に適合する学習機能をもった構文解析技術 (Kazama 2002; Yoshiaga 2001; Miyao 2002)
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

[†]CREST, Japan Science and Technology Corporation
[‡]Graduate School of Information Science and Technology, University of Tokyo
[§]http://www.kototoi.org/

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HPSG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。

2. 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）

(a) 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 (Takahashi 2002)
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。

3. 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）

(a) 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フレームを抽出する技術を開発した。

(b) 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。

4. 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）

(a) 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。

(b) オントロジーを利用した文書検索 (Ogure 2001)
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。

4. 今後の展望

プロジェクト初年度は、基盤となるツール、理論の整備とともに、4つのグループ間の研究交流と共同研究を行うための方向性の議論を集中的に行った。今後は、初年度の基盤整備をさらに進め、4つのグループ間での研究の共同化を具体的に進める。

5. 研究発表

J. Kazama, T. Makino, Y. Ohta and J. Tsujii: Tuning Support Vector Machines for Biomedical Named Entity Recognition, in Proc. of the Natural Language Processing in the Biomedical Domain (ACL 2002) , Philadelphia, Jul., 2002

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, Mar., 2002.

T. Ninomiya, T. Makino and J. Tsujii: An Indexing Scheme for Typed Feature Structures, in Proc. of COLING2002, Aug., 2002 (to appear)

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.

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

T. Takahashi, H. Soonsang, K. Taura and A. Yonezawa: World Wide Web Crawler, the Eleventh International World Wide Web Conference, 2002.

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, Sep., 2001.

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, Jul., 2001.