

Research of Natural Language Processing and Understanding 自然言語の処理と理解に関する研究

プロジェクトリーダー 辻 井 潤 一

東京大学 大学院理学系研究科 教授



1. 研究の目的

WWW、Inter/Intra-Netの普及に伴って、電子化され、自由にアクセスできる自然言語テキストの量は急速に増大しています。また、辞書、シソーラスなど、自然言語処理に必要な言語資源の電子化も急速に進展してきています。しかしながら、情報検索・テキストからの情報抽出・知識獲得といった大量の現実テキストの集合を対象とする応用システムでは、言語の構造的な側面に焦点をあてた自然言語処理・理解研究の成果はあまり使われていません。むしろ、言語の統計的な側面からの手法が主流となっており、言語の構造的な側面をとり扱う場合にも、応用分野に依存したアドホックな手法を用いるに留まっています。

本研究は、(1) 計算言語学、言語理解研究からの手法を現実の大規模なテキスト集合に適用する際の問題点を明らかにし、その解決手法を提示すること、(2) これまで独立に研究されてきた「構造を中心とした」手法と「統計的な側面からの」手法とを統合する枠組を提示すること、および、(3) (1)(2)の成果を各種の応用システム(知的な情報検索、テキストからの情報抽出・知識獲得など)を構成することで検証すること、を目的としています。

2. 研究の内容

言語の構造的な側面を中心にした処理といった場合、まず言語あるいはテキストに関連する構造とは何かを明確にする必要があります。この構造として、我々は、1) テキストの意味構造 2) 分野知識の構造 3) テキスト内 / テキスト間の構造を考えることにします。以上3つの、言語テキストに関連した構造を情報検索に反映するためには、そのための要素技術をまず確立し、それを目的に応じて統合してゆくことが必要です。このため、本プロジェクトでは次の4つの研究を行なっています。

1) 高効率、高耐性で、かつ、言語理論的にも健全な文解析手法の確立

現時点の言語理論の観点からも妥当で、かつ、効率、耐性の観点からも実用的な文解析手法を確立します。文解析技術は、自然言語処理のもっとも基盤的な技術であり、この技術の確立は、本プロジェクトの今後の研究方向を決定するものです。現在、(a) コンパイル・データフロー解析による HPSG の高速文解析手法の確立 (b) タイプつき素性構造抽象機械の実現とそれに基づく論理型プログラム言語 (LiLFeS) の開発 (c) 並列計算による HPSG 文解析の高速化 (d) ペンシルバニア大学の LTAG 文法の HPSG への移行と大規模な英語文法の開発、のように有望な結果をあげています。

2) 言語、および、実世界知識の獲得手法の確立

実世界知識の言語的な反映としてのシソーラスを自動的に獲得する手法の確立は、分散的で多様化したネットワーク時代の情報検索にとって不可欠の技術です。また、シソーラスだけでなく、分野やテキストタイプごとに变化する制約を自己組織的に獲得する技術は、将来の自然言語処理技術の鍵となる技術であり、同時に、人間の言語獲得・知識獲得のモデル化の第一歩となるものです。

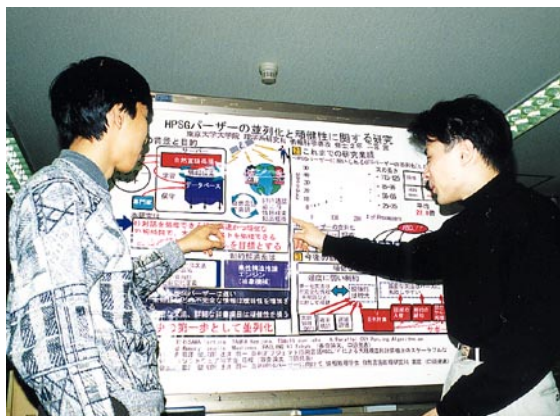
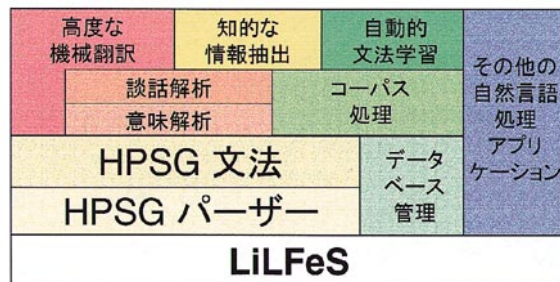
本プロジェクトでは、(a) 文法モデルと統計的な言語モデルの統合を行なう実験 (b) 専門用語認定のアルゴリズム開発と医学コーパス (眼科病理) への適用 (c) 論理型言語 LiLFeS とコーパス・データベースとの結合に関する基本設計と開発 (d) 周辺の言語構造から専門用語の意味的クラスを推定する計算モデルの作成、のように知識獲得を行なうための統合システムの設計開発と個々の知識獲得手法の研究とを同時並行的に行なっています。

3) テキスト内 / テキスト間構造の処理モデルの確立

テキスト間に見られる相互関係の認識の第一歩として、テキスト・クラスタリングの研究、テキスト内の照応関係・結束構造の解析に関する研究を行なっています。テキスト・クラスタリングの研究では、大量のテキストを高速にクラスタリングするアルゴリズムを開発しました。テキスト内の照応関係・結束構造の解析に関する研究では、エキスパートシステムの枠組による解析システムを開発しました。さらに本プロジェクトでは、これらの研究を進めるとともに、テキスト・クラスタリングから、テキスト間の内容の重複、時系列関係など、より構造的な側面に焦点を当てたテキスト・スレディングに関する研究、および、相互参照関係や修辞関係を取り扱うための形式性の高い枠組 (たとえば、DRT 理論) の研究を行なう予定です。

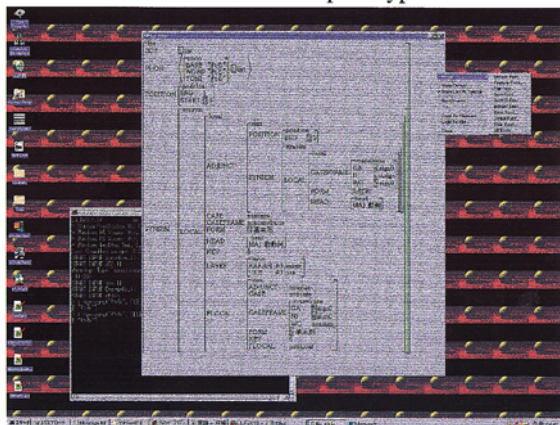
4) 統合化された知的検索システムのプロトタイプ・システムの開発

上記の要素技術の現実場面での有効性を示すために、知的検索システムのプロトタイプを作成します。この研究課題は、プロジェクトの後半、3年度から本格化する予定で、現在はプロトタイプ・システムとして開発する対象分野を調査している段階です。現時点では、(a) ゲノム文献・データベース (b) 東大病院における退院サマリ (c) 特定の話題に関する新聞記事、電子メールでの議論、のような応用分野を想定し、プロジェクト後半からの大規模な実験に備えて、テキスト収集、システム仕様の検討を行なっています。



LiLFeS GUI

Screenshot of the prototype



3. 研究の体制

期 間 1996年10月 - 2001年3月

構 成 プロジェクトリーダー1名、コアメンバー4名、研究協力者16名 (うち日本学術振興会研究員2名)

実施場所 本研究は東京大学・辻井研究室、東京工業大学・徳永研究室、京都大学・黒橋研究室の共同研究として進められています。また、研究を効率よく推進するために国内・国外の研究グループとの研究交流を積極的に進めています。同時に、日本電子化辞書 (EDR) との研究協力のもとに、民間企業の研究グループとも技術の交流を行なっています。