研究課題別事後評価結果

1．研究課題名 「情報のモビリティを高めるための基盤技術」

2．研究代表者名及び主たる研究参加者名

研究代表者

辻井　潤一　（東京大学大学院情報学環　教授）

主たる共同研究者

米澤　明憲　（東京大学大学院情報理工学系研究科　教授）

中川　裕志　（東京大学情報基盤センター　教授）

3．研究内容及び成果

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した研究である。

前期は研究を特定応用のためのアドホックな技術開発にしないために、(1)理論的に健全な枠組に基づく汎用の言語処理技術、(2)本格的なオントロジー処理のための大規模リソースの構築、(3)大量かつ複雑な処理を行うための計算機インフラ技術、(4)ユーザ指向の情報交換を目指すエージェント技術を目指し、4つのグループ（言語処理G、オントロジーG、ソフトウェアG、エージェントG）がそれぞれを担当した。

後期には明確な情報要求をもつユーザ集団（生命科学の研究者）と緊密な研究協力を行うことで、現実的な場面で実際に有効に使用できる統合システムの開発を進めた。結果として(1)深い意味構造を出力する文解析器、(2)テキスト構造と意味に基づく知的な索引構造、(3)機械学習と記号処理の融合のためのアルゴリズム、(4)意味処理・知識処理研究のリソース構築、(5)テキストの収集と処理のためのソフトウェア基盤、(6)ユーザ指向のHCI（Human Computer Interaction）、の分野において著しい成果をあげた。

前記の要素技術開発における主な研究成果を以下に示す。

(1) HPSG(Head-driven Phrase Structure Grammar)による英文解析システム（Enju）：言語学的に妥当な文法に対する確率モデルの定義に成功し、制約を満す結果をすべて出力する能力を持つだけでなく、コーパスから獲得された確率モデルに従って、おのおの結果の確率値も同時に付与することができる。また、コーパス指向の文法開発を提唱し、Genericな言語モデルを現実のコーパスに合わせて精緻化する技術、分野固有の文法制約を学習する手法を確立した。これらの理論を実装した文解析システムEnjuは深層の意味構造を計算するDeep Parser（深層構文解析システム）であり、かつ大規模テキストを現実的な時間で解析できる世界最初のものである。文法の精緻化により、精度の点でも現在世界有数のものとなっている。

(2) 高効率な領域代数処理システム、素性構造データベースの開発：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し、(1)の処理結果を索引構造に反映させることで、指定の意味構造検索を高速に検索するシステムを開発した。また、これを使った統合サービスシステムとして14億語の解析ずみMedlineでの検索システムを構築し、現在EBI(European Bioinformatics Institute)で運用されている同種のシステムより優れた検索能力を持つことを確認した。

(3) 分野に依存したPOSタガー（part-of-speech tagger）、NER（Named Entity Recognition）の開発：MEMM（Maximum Entropy Markov Models）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効果をもち、かつはるかに高速な実装が可能なこと、英語品詞付与やNERに有効であることを実証した。さらに、この同じ手法を英語の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。これらの成果は個別のタスクごとに工夫されてきた機械学習と処理アルゴリズムの関係をタスクに独立な一般的な枠組で統合するものであり、国際的にも高い評価を得ている。

(4) 生命科学の意味つきコーパスとオントロジー（GENIAコーパス）：2000抄録（2万文、50万語）のAnnotated Text（GENIAコーパス）を作成し、世界に公開した。現在240を越える研究チームがこのデータを使って研究している。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・スタンダードとしても活用されている。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ことから、急速に展開しつつある言語的意味と知識の研究においても重要な基礎資料となっている。

(5) テキスト収集の高速クローラーと分散処理のための基盤ソフトウェアGPX：高速クローラー(crawler)の研究では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、そのクローラーを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。また、日常的に利用できるPCクラスタを簡便に使用するためのツールGPXを開発し、実際のテキスト処理に適用することでその有効性を確認した。

(6) ユーザ指向のHCI：言語的・テキスト的な媒体による意思の伝達の対極として、人間とロボットとのインタラクションを例に身体性を考慮した情報伝達を研究した。とくに、引き込み原理と確率推論を使った人間・人工物インタラクション方式を開発し、それを実際のロボットに実装することでその有効性を確認した。引き込み原理を使ったインタラクション方式は、個々のジェスチャの意味づけを予め行わない非記号的なジェスチャによるインタラクションを世界に先駆けて実現したものである。

プロジェクト後半の2年間は、これらの要素技術だけでなく、それらを統合する研究を積極的に推進し、統合的な実験や統合サービスシステムの構築を行った。以下の統合実験は本研究の技

術が国際的な研究水準をはるかに凌駕するものであることを実証した。また、多様な実ユーザを対象にしたサービスシステムを構築することで、「Genericな言語処理に基づくアドッホクでない基盤技術であり、かつ実ユーザの情報要求に応えるための基盤技術」が可能なことを実証した。

(7) 統合実験：分散計算環境のツールGPX、HPSGによる文解析器(Enju)、機械学習によるPOS/NER、GENIAコーパスという、4つの研究成果を統合することにより、生命科学分野のテキストベースMedline抄録全体（14億語）を8日間で処理することに成功した。これは深い文解析を使った実験としては従来の研究を質・量ともに大きく凌駕するものである。分散処理を行わない場合には、この種のシステムでは最速のEnjuを使っても2年以上の時間を要する。

(8) 統合的なサービスシステム：明確な情報要求を持ったユーザ（生命科学分野の研究者）と緊密な共同研究を行い、本プロジェクトの成果をかれらのためのサービスシステムとして実現することで研究の有効性を示した。具体的には、病疾患・遺伝子の関係発見を援助するシステム、蛋白質相互作用の抽出システム、Medlineの知的検索システムを作成した。また、ユーザを特定しない専門用語認識システム(言選)や多言語情報検索システム(KIWI)を開発、公開した。

これらの要素的な研究成果の統合により、構造的に複雑な言語処理・知識処理を大規模なテキストベースに対して適用できる基盤技術を開発すると共に、実際にその有効性を実験により確認した。研究の最終的な成果は性質の異なるユーザ集団を対象とした複数の統合的サービスシステムの形態でユーザに公開されている。

このような複数分野の最新成果をテキスト情報の収集・処理・提供のために統合して、系統的な基盤技術を開発する研究は本研究開始時の5年前には世界でも全く行われていないものであった。

4. 事後評価結果

4－1. 外部発表(論文、口頭発表等)、特許、研究を通じての新たな知見の取得等の研究成果の状況

ネットワーク上に存在する膨大なテキスト情報を効率的に利用するために、言語処理、知識処理、GRIDコンピューティング技術、エージェント技術の研究成果を統合することによって、膨大なテキスト情報の収集と処理、ユーザへの情報提供に関する基盤技術を確立した。改良された文解析システムは強力であり、言語処理研究者に広く使われるであろうし、生命科学分野における意味付きコーパスであるGENIAコーパスの規模と質は非常に優れたものである。この基盤技術を生命科学分野に適用し、14億語という膨大なテキスト処理の統合実験を世界に先駆けて行い、処理速度、精度、質、量共に世界最高レベルの能力を持つことが実証されたことは高く評価できる。世界の有力な研究グループがシステム構築のために利用し、かつ多くの論文に引用されていることは国際的な評価も高く、この分野における研究をリードしている証拠といえる。

実証システムの応用対象として生命科学分野の英文テキストが選定されたが、生命科学以外の分野に同様の手法が使えるかどうかはこれからの課題である。今後、研究成果の水平展開を図れば適用分野は広がる可能性は高く、そのための努力を期待したい。

論文発表は国内14件、海外35件、口頭発表は国内49件、海外78件である。この研究は特許な

どの工業所有権にはなじまないものであるが、出願が全くなかったことには不満が残る。しかしながら、知的財産としては価値のあるものを多く作り、全てのものを公開という原則でこの分野の世界の研究者のために貢献しているのは素晴らしいことである。

4－2. 成果の戦略目標・科学技術への貢献

本研究はいわゆるソリューション型ではなく、理論面を深め要素技術を積み上げて行ったものであり、今後の研究展開の基礎と基盤を提供するものといえる。更には、より高度な推論を含む検索など新たな研究テーマが明らかになってきており、そのような新しい研究の展開が期待される。

具体的に研究成果を適用した生命科学分野においては大きなインパクトを与えると共に、この分野の多くの研究者に利用され、研究の進展に貢献することが期待される。更に適用分野を広げることによって幅広い科学技術の発展への貢献が可能となるであろう。

4－3. その他の特記事項（受賞歴など）

国内3件、国際4件の受賞があるが、研究代表者が多くの国際会議において基調講演、招待講演に招かれていることや、英国マンチェスター大学の国立テキストマイニングセンターへ招聘されたことなどは、研究成果が国際的に高く評価されている証左として特筆すべき点である。