自然言語の処理と理解に関する研究

Research of Natural Language Processing and Understanding

（研究プロジェクト番号：JSPS-RFTF 96P00502）

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
コアメンバー
佐藤　泰介　　東京工業大学大学院工学系研究科・教授
大江　和彦　　東京大学大学院医学系研究科・教授
黒橋　禎夫　　京都大学大学院工学系研究科・講師
徳永　健伸　　東京工業大学大学院工学系研究科・助教授
鳥澤健太郎　　東京大学大学院理学系研究科・助手
白井　清昭　　東京工業大学大学院工学系研究科・助手

１．研究目的

(1) 言語学的に洗練された文法的枠組による処理と実用的言語処理とのギャップを埋める技術の構築
(2) 構造中心の言語処理と大規模コーパスにもとづく確率・統計処理を融合する基礎理論の構築
(3) 知的な情報検索を実現するための言語処理技術の開発
(4) 知識処理と言語処理とを融合する技術、特に、一文を超えた文脈処理モデルの構築

以上の４つは、現時点での言語処理の要素技術に対するブレークスルーを目指すものである。これらのブレークスルーの応用面へのインパクトを実証するために、

(5) 実の要求が明確な適用分野を選定し、開発した言語処理技術がその要求にどのように応えられるかを具体的なシステム（複数）の構築を通して明らかにする。

２．研究成果概要

2.1　言語学的に洗練された文法的枠組による処理と実用的言語処理とのギャップを埋める技術の構築

言語学的に洗練された文法枠組みを実用的な処理に適用する際の最大の技術的ギャップは、その処理速度にある。本プロジェクトでは、処理速度の向上のために、（A）型付素性構造を処理する論理型言語（LiLFeS）の開発、（B）文法記述のコンパイルと多段階処理、（C）曖昧さのパッキング手法という 3 つの独創的なアイデアに基づき、従来手法に比べて数百倍の処理速度の向上を得た。この速度向上が大規模文法においても有効であることを、米国スタンフォード大学・ペンシルベニア大学、ドイツ・ザールブリュッケン大学と共同で開発した大規模文法に適用して実証した。このシステムは、現時点で世界最高の処理速度を持つ。

2.2　構造中心の言語処理と大規模コーパスにもとづく確率・統計処理を融合する基礎理論の構築

2.1 で開発した処理システムと、統計モデルである ME（Maximum Entropy）、決定木、SVM を融合することで、文節係り受けの解析において約９０％の解析精度が得られることを示した。また、浅い構造処理と HMM との組み合わせが、特定分野の専門用語の自動認識や、意味クラスタの自動構成に有効であることを示した。

2.3　知的な情報検索を実現するための言語処理技術の開発

文解析結果を利用した索引語の自動付与、および、テキスト中の単語共起関係からのシソーラス自動構築の手法を開発し、その有効性を知的検索システムに組み込み実証した。このシステムは、米国での TREC コンペティションで 56 チームのシステム（その多くは商用システム）と競合し、4 位の好成績を上げた。

2.4　知識処理と言語処理とを融合する技術、特に、一文を超えた文脈処理モデルの構築

テキスト情報をそのまま知識として利用する枠組みを開発し、その有効性の検証として、辞書定義文から日本語助詞「の」の意味解釈、名詞句間の関係認定を行う文脈モデルを開発した。この枠組みは、知識と文脈処理の新方式として注目され、民間企業との技術提携の議論が進んでいる。また、テキストからのオントロジー構築技術を開発したが、この技術は、現在、ゲノム科学でのオントロジー構築の研究に発展している。

2.5　実の要求が明確な適用分野を選定し、開発した言語処理技術がその要求にどのように応えられるかを具体的なシステム（複数）の構築を通して明らかにする

テキストを知識とする対話システム、ゲノム分野の情報抽出システム、医学分野での知的情報検索システムを構築し、本研究での要素技術の有効性を確認した。

３．結論

本研究の成果によって、言語をとりまく周辺科学の知見を、処理のための「計算理論」へと統合していく鳥瞰的な見通しが得られた。科学

的な方向としては、この鳥瞰図のもとに、心理言語学・大脳科学・理論言語学との積極的な共同研究を進めることで、人間の言語処理機構の詳細なモデルを構築していくことが重要である。この種の基礎的研究は、人間にやさしいインターフェースをもった情報機器を構築する、真に革新的な理論作りに貢献しよう。

より短期的には、本研究で構築した言語処理技術、テキスト管理技術、知識と言語の処理モデルは、ネットワークを介した知識共有システムを構築する基盤技術となる。今後は、これを実際の大規模な知識やテキストをもった分野に適用し、その有効性を確認すること、また、その過程で未解決な技術課題を定式化し、解決していくことが重要である。

本研究は、言語処理の現時点での到達点を明確にした。もちろん、言語処理がカバーする範囲は広い。言語的と非言語的な情報表現（画像、グラフなど）の相互関係、言語が使用される非言語的な状況と解釈の問題、音声言語との関連などは扱えなかった。このような言語と非言語の問題を、本研究で構築した言語プロパーの枠組みの中にどう取り込めるかは、今後の課題である。

主な発表論文

(1) Yusuke Miyao, Takaki Makino, Kentaro Torisawa and Jun'ichi Tsujii, “The LiLFeS abstract machine and its evaluation with the LinGO grammar,” Journal of Natural Language Engineering, Cambridge University Press, **6 (1)** (2000), 47-62

(2) Kentaro Torisawa, Kenji Nishida, Yusuke Miyao and Jun'ichi Tsujii, “An HPSG parser with CFG Filtering,” Journal of Natural Language Engineering, Cambridge University Press, **6 (1)** (2000) 63-80

(3) Hideki Mima and Sophia Ananiadou, “An application and evaluation of the C/NC-value approach for the automatic term recognition of multi-word units in Japanese,” Terminology, **6 (2)** (2001) 175-194

(4) Sadao Kurohashi, and Yasuyuki Sakai, “Semantic Analysis of Japanese Noun Phrases: A New Approach to Dictionary-Based Understanding,” Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL99), (1999) 481-488

(5) Rila Mandala, Takenobu Tokunaga and Hozumi Tanaka, “Query expansion using heterogeneous thesauri,” Information Processing and Management, **36-3** (2000) 361-378