NICT 自然言語グループの紹介

情報通信研究機構　けいはんな情報通信融合研究センター
自然言語グループリーダー兼タイ自然言語ラボラトリー長
井佐原　均

情報通信研究機構（NICT）の自然言語グループでは、自然言語処理技術の研究開発を行っており、その基盤として、言語に関する研究と、研究用言語資源の開発を行っている。また、アジア圏における研究協力の拠点として、バンコク北部にタイ自然言語ラボラトリーを設置し、アジア言語の研究開発を行っている。さらに、けいはんな情報通信融合研究センター内にオープンラボを設置し、企業との連携の下で、自然言語処理システムに関する研究開発を行っている。以下、これらの各項目について概説する。

自然言語処理技術の研究開発

自然言語グループでは主としてコーパスからの学習に基づく手法を用いて、自然言語処理技術の研究開発を行っている。研究対象としては、文解析・生成といった汎用的基盤技術から、情報検索・抽出、要約・言い換えといった応用技術まで、幅広い研究を行っている。また、新しいテーマとして、教育支援システムや機械翻訳システムの研究も開始した。

文解析においては、学習に基づく高精度の形態素解析システムを開発した。このシステムは先で述べる日本語話し言葉コーパスへの形態素情報の付与に用いられ、高精度の言語データの効率よい作成に寄与した。また、2 言語の文を入力とする第 3 言語翻訳方式による機械翻訳システムや大規模な学習者コーパスを利用した学習支援システムの研究開発などユニークな研究開発も行っている。

言語に関する研究

自然言語処理の基盤となる、言語そのものに関する研究として、語の意味を客観的に表現する語彙意味論の研究、談話理解や感性情報処理の研究、意図の抽出の研究を行っている。

語彙意味論の研究においては、語の意味を 2 次元平面上に表す意味マップを拡張することにより、語義の階層構造を実際の言語データから客観的に求める手法を開発した。談話理解においては、3 人対話の収集と解析など、新しい観点からの研究を行っている。感性情報処理の研究においては、被験者実験に基づく音楽感性や敬語運用に関する研究を行っている。意図の抽出においては、アンケートの自由回答を分析し、分類する手法の研究を行った。

研究用言語資源の開発

研究用言語資源とは、文法や辞書、大量の文章（コーパス）といった言語研究で用いるデータやツールを指す。我々は、約 1300 人の日本人の英語発話を収集した世界最大の英語学習者コーパス（NICT JLE Corpus）や、日英中の対訳データを中心とし、様々な言語情報を付与した大規模なコーパスである NICT Corpus、国立国語研究所と共同で作成した日本語話し言葉コーパス（Corpus of Spontaneous Japanese）を開発してきた。また、大規模電子化辞書であるEDR電子化辞書の全著作権を所有し、改良に務めると共に、一般に公開している。

タイ自然言語ラボラトリー

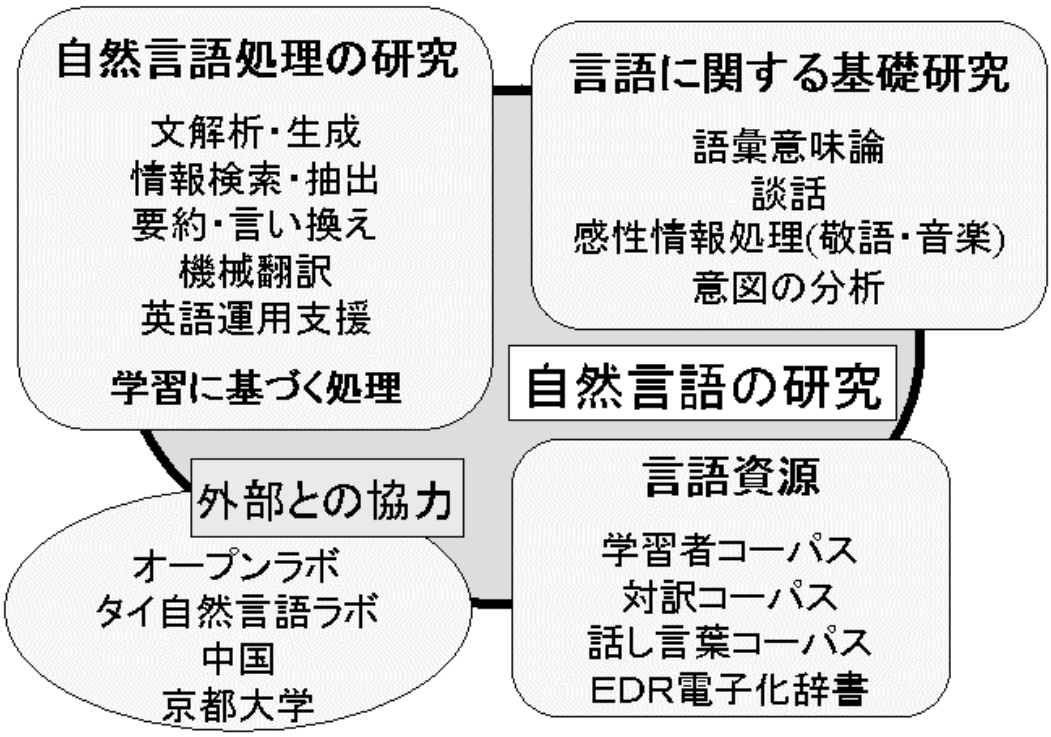
NICTでは、タイにアジア研究連携センターを設立し、その中にタイ自然言語ラボラトリーを設置している。ここでは、タイ及びバングラディシュからの研究者により、アジア言語の辞書の研究開発、電子図書館の研究開発、文書フォーマットの標準化、個人情報管理システムの開発を行っている。また、ベトナムやミャンマー等の近隣諸国との交流も進めている。

今後さらに本ラボラトリーを充実させ、アジア圏における自然言語研究の拠点として、活用する予定である。

けいはんなオープンラボ

自然言語処理に興味を持つ企業のとの共同研究として、オープンラボにおいて、自然言語処理応用技術の研究開発を行っている。具体的には、対訳コーパスの開発、インターネットからの情報抽出の研究開発、英文読解支援や機械翻訳用知識の自動獲得の研究、ウェブからの情報獲得を容易にするインタフェースの研究等を行っている。

オープンラボにおける研究は、従来のNICT単独による研究開発よりも、さらに実用化に近い観点からの研究開発を行っている。



自然言語グループの概要