3.1.6　自然言語グループ／タイ自然言語ラボラトリー

中期計画期間全体

目　　標

人間の知的活動を支援する環境の実現のため、人間のコミュニケーションの基幹をなす自然言語の処理と伝達の技術を開発する。中期目標期間終了時点では、言語情報を多言語かつ多様な形態で処理するシステムを開発する。

目標を達成するための内容と方法

システム開発を視野に入れた学習に基づく自然言語処理の研究を中心に、ブレークスルーを目指す自然言語の基礎研究にも力を注いでいる。また、開放的融合研究を行うなど、他機関との協調、競争的資金の獲得にも努力している。

特　　徴

自然言語処理技術において、コンテスト等によって最高水準の成果を客観的に示している。また、第三言語翻訳などの新しい手法も提案している。自然言語の基礎研究を合わせて行うことにより、他の研究機関にない幅の広い研究活動が可能となる。

今年度の計画及び報告

今年度の計画

自然言語処理の研究では、学習に基づく言語理解・言語生成・言語分析・応用の研究を行う。多言語処理を念頭に、2言語解析とキーワードによる文生成の研究を行う。自然言語の基礎研究では、語彙意味論、対話モデル、感性情報処理の研究を行う。開放的融合研究では、最終年度に向けてプロトタイプの要素技術を開発する。

今年度の成果

⑴　自然言語処理の研究においては、学習に基づく処理技術の研究を進め、言語理解、言語生成、情報検索等の要素技術を開発した。第三言語翻訳システムの要素技術となる解析技術、生成技術の研究を行った。情報検索、質問応答に関しては、評価コンテスト型のワークショップに参加し、最高レベルの成績を上げた。コーパス（新聞記事等を題材に、その文章をデータ化したもの）を用いた分類手法の研究を進め、学習者コーパスを用いて、学習者の能力を判定する実験を行い、高い精度を得た。

⑵　自然言語の基礎研究においては、神経回路網、補完類似度等の手法を大規模コーパスに適用し、語彙の形式化を目指した。語の意味の階層構造を客観的に表現できる枠組みを開発した。談話（文脈）解析の研究を行った。これに沿って、下に述べる話し言葉コーパスに対し、談話タグを付与している。1200名（300時間）分の日本人の英語発話を収集し、書き起こすことにより、世界最大級の英語学習者コーパスを開発した。このコーパスには、書き起こしに関する情報と、発話者の誤りに関する情報が付与されている。

⑶　開放的融合研究においては、話し言葉を高精度に解析するシステムを開発した。このシステムを用いて、話し言葉コーパス全体（700時間分の講演書き起こし）に対して、形態素情報を付与する。付与の精度を上げるために、人手修正を行っているが、このシステムの特性により、効率の良い人手修正が可能となった。形態素情報に加えて、係り受け構造についても情報付与を行っている。また、話し言葉要約システムプロトタイプの開発のため、要約システムの研究開発を行った。要約システム開発の基礎データとして、コーパスから重要文を抽出したデータとコーパスを基に自由に要約を行ったデータを作成している。

⑷　共生コミュニケーションシステムにおいては、実用システムの開発の一環として、質問応答システムを開発するとともに、機械翻訳及び英文読解に関して、民間企業の研究員を取り込んだ研究開発を開始した。

⑸　多言語処理に関しては、CRLアジア研究連携センター内にタイ自然言語ラボラトリを設置した。現地の研究者を雇用し、アジア圏におけるオープンソースソフトウェアの開発と普及及びアジア圏言語の多言語辞書の開発を開始した。さらに、中国語処理の研究開発を開始した。日中機械翻訳システムの開発に向けて、日本語新聞記事を翻訳することにより、日中対訳コーパスを作成している。このデータを用いて、日中対訳辞書の開発を開始した。また、自己組織化意味マップを中国語に適用し、新たな実用を目指している。

⑹　言語横断文検索においては、大規模な対訳データとコーパス検索ツールの公開を行った。読売新聞記事とデイリー読売（英字）新聞記事とを自動的に対応付け、また、記事中の文間にも対応を付ける技術を開発し、対応付きデータを作成し、研究用に一般公開した。このデータは現在利用可能な対訳データとしては最大規模のものである。