セマンティックアーカイブコア客員プロジェクト研究報告

増山繁, 中川聖一, 秋葉友良 (セマンティックアーカイブコア)

セマンティックアーカイブコアでは、客員教員として長岡技術科学大学山本和英准教授、名古屋大学北岡教英准教授、法政大学伊藤克亘教授を迎え、講義支援システムにおける Web からの関連情報抽出、講義音声の自動要約、特許文の機械翻訳について研究を行った。

(山本和英客員准教授・増山繁教授)

セマンティックアーカイブコアでは、講義収録データに基づき講義コンテンツを蓄積するとともに，その要約，インデキシングなどの自動処理を行い，講義教材をより高度に利用できるシステムを構築している．しかしながら、学習者にとって必要な情報が講義コンテンツからのみ得られるとは限らない．一方，情報爆発の時代といわれるように，Web には大量の情報が溢れており、その中には学習者が必要な講義補完情報が含まれている可能性が高い．そこで、講義への補完情報の Web からの取得をおこなう研究を行なった．具体的には，テキスト化された講義内容と類似した外部 Web ページを検索する手法についての提案と意見交換を行った。この手法はテキスト内の共起情報をグラフ化し、各語句の貢献度を計算することで講義コンテンツ中の重要語句を同定し、この結果に基づいてサーチエンジンで Web ページ検索を行う。提案手法を実装した結果、比較的長いテキストを対象とした場合に本手法の有効性を確認した.

(北岡教英客員准教授・中川聖一教授)

ネットワークの容量増加に伴い、講義や講演をビデオ録画し、自宅からでも容易に学習や復習が可能なシステムが実用化されている。また、講義中に使用しているスライドの切り替わりとビデオを自動対応付けするツールもある。しかし、このようなシステムやツールには動画の再生、早送り、巻き戻しや、スライドバーによる任意位置からの再生など、標準的な閲覧環境しか実装されていない。e-Learning の長所はいつでも好きな時に学習、復習が可能なことであるが、そのビデオは基本的に通常速度で再生されるため、利用には収録時間と等しい時間が必要となる。　　中川研究室では、講義の際に収録する音声や動画といったメディアファイルに対して、書き起こし、要約、セグメンテーションやインデキシングを自動的に行い、効率的な学習を支援するシステムを構築し、被験者実験にて有効性を確認している。本客員教員プロジェクトでは、音声要約に関して共同研究を行った。我々は従来から重要文抽出に基づく要約手法を検討してきたが、人間による要約では、連続して重要文を抽出する場合が多いという知見を、今回従来手法に組み込み、評価実験により有効性を示した。

(伊藤克亘客員教授・秋葉友良准教授)

インターネットをはじめとした世界規模の情報ネットワークの発展により、異なる言語を母国語とする人間の間でコミュニケーションを行う機会が増大しつつある。機械翻訳は、このような状況を支援する基盤技術として今後ますます重要となる。本プロジェクトでは、特許文翻訳を対象に、翻訳例を元に自動的に学習した翻訳知識を元に機械翻訳を行う統計的機械翻訳の研究を行った。対訳データの分野依存性が単語翻訳モデル学習に与える影響を調査し、分野毎に翻訳モデルを構築することによる翻訳の品質への効果を調べた。また、このアイデアを文書の自動クラスタリング技術と組合せることにより、対訳データを自動的に分類し、分野依存翻訳モデルを自動構築する手法を構築した。この枠組みを利用して、機械翻訳の評価型ワークショップ NTCIR PATENT-MT タスク (国立情報学研究所主催) に参加し、評価を行っているところである。