# セマンテイックアーカイブコア客員教授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）

**（伊藤克亘客員教授・秋葉友良准教授）**

個人の生活や体験を様々なセンサ(カメラ、マイク、GPS など)を用いて記録した「ライフログ」を、備忘録や自動の日記作成、業務内容の共有などへ利用することが期待されている。ライフログはデータ量が膨大かつ冗長で、効率的な利用のためには検索のためのインデキシングや要約が必要である。そこで、音響ライフログに含まれる 5 秒程度のイベントのクラスタリング手法について検討した。まず、音響ライフログから冗長部分を除去するために音響情報を約 5 秒のセグメントに分割し、スペクトルと振幅を利用して冗長な部分である無音セグメントの除去を行った。そして、残ったセグメントに対して k-mean クラスタリングを適用し分類を行う。このクラスタリングのための効果的な特徴量を調査したところ、従来、長い(1 分以上)セグメントのクラスタリングに有効であると知られているスペクトル包絡以外に、スペクトルや振幅の差分値で良好な結果が得られることが明らかになった。

**（山本和英客員准教授・増山繁教授）**

今年度は組織にアーカイブされた電子文書の高度利用のための検討を行った。現在、企業において社員の勤務状況を勤務日報という形で電子的に報告させ、管理している。日誌を閲覧する際、直接的に報告されている障害に対しての対策は容易であるが、隠れた障害や次に起こり得る障害を人間が発見し対策することは極めて困難である。隠れた障害を見落とし、対応が遅れてしまうことによって企業の信用を失う事態に発展してしまう可能性がある。もし次に起こりうる障害を予知し、注意を喚起できるならば日報の閲覧コストを下げると共に障害が表面化する前に対処することが可能となる。我々は日報を入力として、次に起こり得る障害を予知するシステムを提案した。評価の結果、ベースラインとして用いたランダム出力の精度を越えることが出来た。また、システムの予知能力は人間の予知能力を越える可能性があることを示した。さらにシステムは人間の気付きにくい障害をある程度予知できることを示した。

**（北岡教英客員准教授・中川聖一教授）**

講義音声に対して高精度な重要文抽出を実現するために、まず、重要文の連続性に着目し、これを動的素性と差分素性として反映させた。つまり、動的素性とは直前の文が重要文として抽出されたかどうかを示す素性、差分素性とは直前の文の素性と現在の文の素性

の差を示し、重要な文が連続すれば、内容が異なっても、差分素性は小さくなる。

次に、内容の重複度合いを示す素性を新たに導入した。冗長性を考慮した素性を導入することで、抽出された文全体に依存関係が生じるために動的計画法では一意に最適解を決定できなくなるので、ビームサーチを導入することで近似解を得た。

これらの素性を従来の素性（表層的言語情報、韻律的特徴）に加えて、重要文の抽出を行った。評価尺度として $\kappa$ 値と Rouge を用いた。4 人の講義音声の書き起こし文に対して、従来法は、それぞれ 0.388 と 0.696 であったのが、連続性の考慮により 0.401 と 0.711 に向上し、さらに冗長性を考慮することにより 0.404 と 0.711 に向上した。一方、音声認識結果文に対しては、それぞれ、0.375 と 0.680, 0.395 と 0.702, 0.391 と 0.699 となり、連続性の考慮が効果あった。人間による要約結果に対しては、$\kappa$ 値が 0.469、Rouge が 0.695 であり、人間並みの要約結果が得られたと言える。

**（梅村祥之客員准教授・増山繁教授）**

我が国は工業社会から知識社会へと急激に移行しつつあり、知財の創出と保護は国をあげての重要事項となっている。通常、特許出願後、特許権成立か不成立かが確定するまでに長年に渡る未確定の時期が続く。そのため、特許調査の担当者は、公開特許公報に基づき特許権成立の可否を推定しなければならない。そこで、本年度の研究では企業の研究開発部門における特許調査の支援を目指して、公開特許公報のテキストデータから、その特許権が将来成立するか否かを推定する手法について基礎検討を行った。特許性を反映し、かつ、機械処理可能な素性として、表層的なテキスト処理で算出する素性を定義した。特許権成立の可否が確定している 35,818 件の公開特許公報のテキストデータを用いて機械判定を行い、ベースラインとしたランダム抽出に対する性能向上を確認した。