地域アーカイブにおけるプライバシ情報マークアップ支援システム

盛永 浩太

デジタルアーカイブとは、人びとのさまざまな情報資産をデジタル媒体で保存し、共有し、活用する仕組みの全体像である。本研究では、地域の画像とそれにまつわる思い出話のテキストデータをアーカイブ対象として取り扱う。

地域アーカイブは地域社会における歴史的、文化的、そして社会的な価値の保存と伝承のために構築されたアーカイブである。本研究では、福島県双葉町を対象として、地域の写真とそれにまつわる思い出話のアーカイブ化に焦点を当てる。

思い出話を地域アーカイブとして記録するにあたって、テキストの中にプライバシ情報が含まれることが考えられる。アーカイブの適切な保存と利用のためには、それらのプライバシ記述を適切に保護することが重要である。ただし、ユーザに一からプライバシ情報を指定してもらうと、ユーザが気軽に思い出話を投稿できないことが考えられる。そこで本研究では、ユーザによるプライバシ記述の指定を支援する目的で、思い出話のアーカイブ化におけるプライバシ情報の秘匿を支援するシステムを構築する。

本システムは、ユーザが記述した思い出話のテキストに対し、プライバシ情報に当たる部分を自動で識別し、マークアップをしてユーザに提案する。そのマークアップされた内容に対して、ユーザに適宜修正と確認を行ってもらい、テキストと秘匿部分の情報をアーカイブに登録をする。

テキスト中のプライバシ記述部分を自動で検出し、マークアップするプロセスでは、日本語形態素解析器を用いて固有表現認識を行う。この処理でシステムは個人名、地名、日付などのプライバシにあたる記述を自動的に特定する。ユーザがチェックと修正を行うプロセスでは、マークアップの追加や削除といった機能をドラッグやクリックで操作できるようにして、マークアップ支援を行う。

このシステムの評価に当たって、10 件のテスト用テキストと正解データを用意し、固有表現認識の再現率、適合率と、修正手法についての評価を行った。結果は、再現率が 72%に対して、適合率が 29%と低い値を記録した。再現率が高く適合率が低いことから、ユーザに求められる操作が削除操作であることが考察される。削除操作の容易さと新規マークアップの作成コストを考慮することで、本システムがユーザの負担を軽減できると考えられる。

今後の研究方針としては、固有表現認識部の精度向上を図るため、抽出アルゴリズムの最適化と、より多様な文脈を含む学習データの収集が重要である。また、固有表現で抽出された客観的なプライバシ情報と、ユーザが指定する秘匿部分の差異を分析することが、主観的なプライバシ情報抽出を実現するための重要な検討課題と考える。

（指導教員　阪口 哲男）