平成27年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	斉 藤　航 佑
卒 業 研 究 題 目	「昭和天皇実録」からの固有表現抽出	

昭和天皇実録 (以下、実録) は、昭和天皇の誕生から崩御、そしてその後の葬儀と陵籍登録までを年代順に記したもので、昭和天皇の死後に宮内庁により編纂され、2014 年 9 月 9 日に公開された。 従来より昭和天皇に関しては、政治への関与が多様な解釈と共に論争の種となってきた。それだけに、宮内庁が確実な資料に基づき編集したとする実録は、社会的な注目を集めている。公開直後には各新聞で特集が組まれ、その中でも毎日新聞は、国策に関わる人物が天皇と 1 対 1 やごく少人数の場で会話したとみられる記述を集計することにより、「天皇と国策」の距離や影響を時代ごとに捉える試みを紙面にしている。例えば、第二次世界大戦前から軍幹部との面会が増加した一方で、1947 年に天皇が象徴とされると面会が減ったことが具体的なデータに基づいて示されている。ただし、この集計は記者 15 人が人手で行っており、特定の肩書を持った人物に限られるなど、天皇に面会した人物が網羅的に集計されているわけではない。

一方で、自然言語処理の技術の一つに、人名・組織名・地名・時間表現などの固有表現をテキストから抽出する技術がある。この技術を実録に適用することにより、前述の人手により抽出したデータ以上に網羅的な人名の抽出ができるため、より詳細に分析できる可能性がある。また、この抽出結果の二次利用により、歴史・政治学分野での研究がさらに発展するものと期待できる。

そこで本研究では、固有表現抽出を利用することにより、実録から人名とそれに付随する肩書を半自動的に抽出した。具体的には、機械学習ツール CRF++を用いて自動抽出を行い、その結果を人手で修正することにより、人物名とそれに付随する肩書を抽出した。また本研究では、人手の修正コストを最小限に抑えるため、人手で修正したデータを CRF++の学習データに適宜加えることを繰り返し、CRF++による抽出性能を徐々に改善するというスパイラルな構築を行った。

抽出対象のテキストデータには、89 年分ある実録を画像データから OCR により 1 年毎にまとめたものを用いた。まず、5 年分のデータに対して抽出したい固有表現部分を人手でマークアップし、それを学習データとして機械学習を行った。この 5 年分には、4,040 種のべ 11,215 個の固有表現がマークアップされた。この初期学習を行った CRF++を用いて別の年のデータに対する自動マークアップを行い、人手による修正、学習、自動マークアップを繰り返し行った。マークアップの修正は名古屋大学大学院法学研究科の技術補佐員 1 名が実施した。

初期学習データ 5 年分を用いたシステムにおけるオープンテストの結果は、精度 92%、再現率 85%程度であった。しかし、抽出したい固有表現に対して正しくマークアップされた学習データが増え、オープンテストの結果は精度 95%、再現率 92%程度まで向上した。現在、実録の 89 年分のデータから、16,077 種のべ 38,406 個の肩書と 23,959 種のべ 40,480 個の人名を自動抽出した。そのうち、50 年分のデータにおける 9,178 種のべ 21,682 個の肩書と 13,638 種のべ 22,879 個の人名については、人手による修正が完了している。

なお、今後も作業は継続され、すべての年のデータに対する正しいマークアップが実施される予定である。また、本研究で構築したデータは、名古屋大学大学院法学研究科における研究に活用される予定である。