平成30年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	河 合 恒 輝
卒業研究題目	ニューラルモデルを用いた 「昭和天皇実録」からの固有表現抽出	

固有表現抽出の必要性が高まっている。固有表現抽出とは、計算機を用いた自然言語処理技術の一つであり、情報を抽出するための一分野として知られており、文中から固有表現などを抽出し、あらかじめ定義された固有表現分類 (人物名、日付、地名など) へと分類することを指す。特に人物名は、単語の結びつきの判断の難易度が高く、抽出する必要性が高い。例を挙げると、裁判の判決文は個人情報保護の観点から名前の部分を匿名化しなければ公開できない。しかし、匿名化には人手がかかり、そのため公開が進んでいない。

本研究では「昭和天皇実録」に着目する。「昭和天皇実録」とは、昭和天皇の誕生 (明治 34 年 4 月 29 日) から崩御 (昭和 64 年 1 月 7 日)、昭和天皇武蔵野陵の陵籍登録 (平成 3 年 3 月 30 日) までの生涯を年代順に記したものである。今回「昭和天皇実録」に着目した理由は、「昭和天皇実録」は、昭和天皇について書かれた文献を参考に書かれており、人名などの表記が昭和の未公開の判決文と共通する部分が多いと考えられ、かつ先行研究において固有表現へのタグ付けが既に行われており、学習データおよびテストデータとして利用可能だったからである。

先行研究では、MeCab を用いて「昭和天皇実録」を形態素解析した文に、CRF(条件付確率場) を用いて系列ラベリングを行うことができる機械学習ツール CRF++を用いることで、固有表現抽出をしているが、性能がまだ十分ではない。とくに匿名化においては、より高い再現率が求められる

そこで本研究では、ニューラルモデルを導入して固有表現を抽出し、先行研究の性能を超えることを目指した。

具体的には、ニューラルモデルを搭載した機械学習ツール anaGo を用いた。anaGo は、ニューラルモデルとして BiLSTM を採用している。BiLSTM は、LSTM を改良したものである。LSTM は、自然言語処理でよく用いられるモデルである RNN の発展形である。文の単語のタグの情報を読み込むとき、少し前の単語のタグの情報までしか記憶できなかった RNN と違い、長い文に対しても記憶ができるように改良されたものが LSTM である。また、RNN は記憶する範囲が大きくなると、過去のどの情報がどれくらい影響を及ぼすかが複雑になり過ぎる。それにより、伝えるべき誤差が消滅するという問題を抱えていたが、LSTM はそれも改良してある。さらに、文を末尾からも読み込むことで、先に出現したタグの情報からだけでなく、後に出現するタグからも情報を得ることができるよう改良され、より精度の高い抽出を可能にしたのが BiLSTM である。

本研究へのモデルへの入力は MeCab で分かち書きされた文であり、anaGo に入力することで、単語ベースの単語分散表現だけでなく、文字ベースの単語分散表現を作成できる。この二つの単語分散表現を連結して新たな単語分散表現を作成することにより、CRF++において単語の情報だけを用いた先行研究よりも性能が高くなると考えられる。

先行研究と同じ、学習データ 40 年分 (69257 文)、とテストデータ 10 年分 (16141 文) を用いて検証を行ったところ、先行研究においては、精度 95.4%、再現率 92.4%であったが、本研究では精度は 96.0%、再現率は 96.0%という結果が得られた。これにより、性能の向上、特に再現率を上げるという目的を達成した。