『日本語歴史コーパス』に対する文脈化単語埋め込み情報付与

浅原正幸，加藤祥

単語を低次元の実数ベクトルで表現する技術である単語埋め込みの研究が工学分野で進められている。2018 年になり 文脈を考慮した文脈化単語埋め込み技術が開発された。文脈化単語埋め込みは，その出現ごとに異なるベクトルを割り当てるため，語義のあいまい性解消にも有効である。同技術の一つ BERT は，単語穴埋め課題と 2 文隣接課題を自然言語処理の事前学習モデルである。日本語では BERT の事前学習モデルとして『国語研日本語ウェブコーパス』を訓練データとした NWJC-BERT が整備された。このモデルは言語研究を目的として整備されており，UniDic-分類語彙表対応表『WLSP2UniDic』に登録されている自立語と UniDic 中の全付属語を，UniDic の語彙素ベースで訓練したものである。これにより，『日本語歴史コーパス』を含めた UniDic 体系で整備されたコーパスに対し，ベクトル表現を悉皆付与できるようになった。しかしながら，BERT の事前学習モデルは，GPU などを搭載した機材を用いることが必要で，個人で利用することは困難である。そこで『日本語歴史コーパス』バージョン 2019.12 の一部に対して，文脈化単語埋め込み情報を付与した(以下『BERTed-CHJ』 と呼ぶ)。さらに文単位の埋め込み情報も整備した。『日本語歴史コーパス』の単語単位の類似度だけでなく，文単位の類似度も評価できるようになった。

一方,『日本語歴史コーパス』の一部に対して『分類語彙表』の分類番号付与が人手により進められている(以下 CHJ-WLSP と呼ぶ)。既に「竹取物語」「土左日記」「徒然草」「方丈記」の作業が完了している。以下では，文脈化単語埋め込みの有効性を検証するために CHJ-WLSP の分類項目に認定される，現在(.1641 関係-時間-現在) 125 語・過去(.1642 関係-時間-過去)79 語・未来(.1643 関係-時間-未来) 29 語，合わせて 233 語のベクトルを BERTed-CHJ から抽出し，可視化した。古典語の時間表現について，分布意味論の観点から多義語の語義の識別可能性を検討したので報告する。さらに文単位のベクトルにより評価した作品間類似度などの調査についても示す。