

単語置換を用いた BERT による慣用句曖昧性解消

Idiom Disambiguation with BERT Using Word Replacement

船藤 裕文, 寺岡 文博

Hirofumi FUNATO, Takehiro TERAOKA

拓殖大学大学院工学研究科

Graduate School of Engineering, Takushoku University

Email: 20m308@st.takushoku-u.ac.jp

あらまし：慣用句には、字義的な意味とは別に慣用的な意味があるため、文脈理解において、これらの曖昧性を解消する必要がある。本研究では、用例文中で慣用句の構成単語を異なる単語で置き換え、BERT のファインチューニングを行うことで、慣用句全体に適用できる汎用的な曖昧性解消モデルを構築した。評価実験として、先行研究手法との比較を行い、モデルの有用性を示した。

キーワード：慣用句、語義曖昧性解消、BERT

1. はじめに

慣用句には慣用的な意味と字義的な意味があり、曖昧性を持つ。例えば、「骨が折れる」という表現は大変だ、困難であるといった慣用的な意味と骨折するという字義的な意味を持ち、文章中でどちらの意味で使用されているかを正しく判断する必要がある。

日本語学習者にとって文章中の表現が慣用句であるかという判断は難しく、判断を自動化することで学習の手助けとなる。

橋本ら⁽¹⁾は曖昧性のある慣用句を含む用例を集めた OpenMWE コーパスを構築している。OpenMWE コーパスは Web 上から収集された用例中の慣用句が慣用的な意味で使用されているか字義的な意味で使用されているかを人手によってアノテーションされたコーパスである。慣用句 146 種、用例数 102,334 文が収録されている。高橋ら⁽²⁾は、この OpenMWE コーパスの用例を学習データとして BERT⁽³⁾に適用した慣用句曖昧性解消手法を提案している。BERT は大量の文書データを利用した事前学習によって文脈を考慮した単語埋め込みを出力できるニューラル言語モデルである。高橋らは、BERT の出力のうち、慣用句を構成する名詞、動詞に対応する単語埋め込みとそれぞれの単語の一般的な用法における単語埋め込みを多層パーセプトロンへの入力とした慣用句判定器を構築している。

BERT を利用した文書分類には事前学習済みの BERT を特徴量抽出器として利用し、BERT から出力された単語埋め込み表現と学習データに付与されたラベルで分類器を学習する特徴量ベースの手法と分類器に加えて BERT のパラメータも更新し、学習を行うファインチューニングを用いた手法の二種類がある。高橋らは前者の特徴量ベースの手法を用いているが、慣用句の曖昧性解消タスクにおいてファインチューニングが有効であるかどうかはわかっていない。その為、本研究では慣用句の曖昧性解消タスクの精度向上に加えて、本タスクにおけるファインチューニングの有効性も調査する。

2. 提案手法

本研究では OpenMWE コーパスの用例を使用して BERT のファインチューニングを行うことで二種類のモデルを構築し、それぞれのモデルを判断する用例文によって切り替えて適用することで用例文に含まれる表現が慣用句として使用されているか否かを判断する手法を提案する。構築するモデルは一部の用例文に単語置換処理を施して学習する単語置換ありモデルと用例文を変更せずにそのまま学習する単語置換なしモデルの二種類である。ここで、単語置換処理とは慣用句部分の名詞、動詞を学習データに含まれるランダムな別の慣用句の名詞、動詞で置換する処理である。この処理によってそれぞれの慣用句固有の特徴への過学習を防ぐことで、慣用句の種類に拘らない汎用的な慣用句曖昧性モデルになるように学習が行われると考えた。

それぞれのモデルの学習の流れを図 1 に示す。図の左側が単語変更なしモデルの学習、右側が単語変更ありモデルの学習の流れであり、点線で囲まれた箇所が学習を行うモデルである。まず、BERT に入力する為のトークン化処理の際に文頭に付与される CLS トークンへの BERT の出力と用例文に付与された慣用的な意味であるか字義的な意味であるかのラベルを使用してそれぞれのラベルに対するスコアを計算する。そして、出力されたスコアをそれぞれのラベルに対する予測確率に変換し、入力のラベルとの交差エントロピーで損失を計算する。

推論の際には、学習データと同じ名詞を使用した慣用句が存在する場合は単語置換ありモデル、学習データに同じ名詞を使用した慣用句が存在しない場合は単語置換なしモデルのどちらかを適用する。これは同じ名詞を持つ複数の慣用句がそれぞれの固有の特徴によって相互に与える影響を単語置換処理によって緩和できると考えた為である。適用したモデルから出力されるスコアによって用例文に含まれる慣用句の意味を二値で出力する。

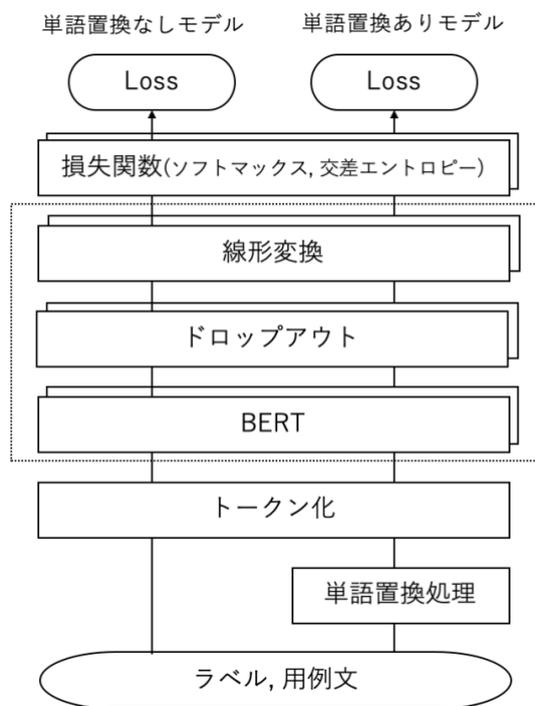


図1 学習の流れ

3. 評価実験

評価データには先行研究に倣い、OpenMWE コーパスの用例のうち慣用的な意味の用例と字義的な意味の用例のどちらも 50 文以上の用例を持つ慣用句 89 種の用例 74067 文を使用した。実験は 10 分割交差検証で行い、学習データに使用した慣用句をテストデータに含まずに検証を行う為、慣用句 89 種を 9 種ずつの 9 組と 8 種の 1 組の 10 組に分け、8 組を学習データ、1 組を検証データ、残りの 1 組をテストデータとした。

BERT のモデルには東北大学が公開している wikipedia の日本語記事で学習された事前学習モデル^{※1}を使用した。また、単語置換ありモデルで単語置換処理を行う割合は慣用的な用例と字義的な用例どちらも 6 割ずつとし、それ以外の用例には変更を加えない。

提案手法との比較の為、高橋ら²⁾の手法で使用されている特徴量を用いたモデルを作成し、提案手法と比較した。また、単語置換なしモデルのみ、単語置換ありモデルのみでも実験を行なった。加えて、高橋らの手法で使用されている多層パーセプトロンに CLS トークンの単語埋め込み表現を入力した特徴量ベースの手法でも実験を行なった。

4. 実験結果と考察

評価実験の結果を表 1 に示す。単語置換なしモデルと単語置換ありモデルを組み合わせた提案手法が最も精度が高くなった。高橋らの手法は論文では正解率 0.836 とされていたが本評価実験ではそれよ

りも低い精度となった。原因として、使用した BERT モデル、形態素解析器等の違いが考えられる。また、単語置換なしモデルと比べ、単語置換ありモデル、提案手法はどちらも精度が高くなった。これは慣用句の曖昧性解消において単語置換処理が有効であることを示している。

また、表 1 における特徴量ベース手法と単語変更なし手法はどちらも CLS トークンに対しての単語埋め込み表現を特徴量として分類器への入力に使用しており、両者の違いはファインチューニングの有無である。本研究の評価実験ではファインチューニングを行なったモデルの方がそうでないモデルに比べ、精度が高くなった。今までの BERT を用いた慣用句曖昧性解消に関する先行研究においては、特徴量ベースの手法が取られていたが、ファインチューニングを行なった方が精度が高くなることが本評価実験で確認できた。

表 1 実験結果

手法	正解率	F 値
高橋ら手法	0.750	0.642
特徴量ベース手法	0.767	0.674
単語置換なしモデルのみ	0.843	0.800
単語置換ありモデルのみ	0.844	0.800
提案手法	0.848	0.806

5. まとめ

本研究では OpenMWE コーパスの用例を使用して BERT のファインチューニングを行うことで構築された単語置換ありモデルと単語置換なしモデルの二種類のモデルを組み合わせて慣用句曖昧性解消を行う手法を提案した。

評価実験により単語置換処理とファインチューニングの有効性を確認した。

謝辞

本研究は JSPS 科研費 JP18K12434 の助成を受けたものです。

参考文献

- (1) 橋本力, 河原大輔: “日本語慣用句コーパスの構築と慣用句曖昧性解消の試み”, 情報処理学会研究報告, 2008-NL-186, pp.1-6, (2017)
- (2) 高橋良輔, 笹野遼平, 武田浩一: “文脈化単語埋め込みを用いた慣用句判定”, 情報処理学会研究報告, 2020-NL-245, pp.1-5, (2020)
- (3) Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186 (2019)

※1 cl-tohoku/bert-base-japanese-whole-word-masking