Vol.2022-NL-253 No.3
2022/9/29

疑似訓練データを用いた BERTによる同形異音語の読み推定

小林 汰一郎[1,a)]　古宮 嘉那子[2,b)]　新納 浩幸[3,c)]

概要：日本語には読みに曖昧性を持つ単語が多数存在する．例えば「辛い」は「カライ」のほかに「ツライ」と読むこともできる．このような単語を同形異音語と呼ぶ．本論文では, BERT を用いて同形異音語の読み推定を行う．訓練・テストデータには現代日本語書き言葉均衡コーパス (BCCWJ) と日本語話し言葉コーパス (CSJ) を利用した．BCCWJ の大半を占める非コアデータの読みは, 形態素解析システム MeCab により機械的に割り振られたものである．また, BCCWJ は書き言葉であり, CSJ は話し言葉なので, ドメインのずれが想定される．CSJ をターゲット領域としたとき, 通常はこの領域の訓練事例を用いて読み推定のモデルを学習・構築すればよいが, 訓練事例の構築コストが高いという問題がある．本研究では自動的に付与されたドメイン外の大量の疑似データ (BCCWJ のデータ) を利用することで, 本来必要としたターゲットの領域の訓練事例の量を大幅に削減することができた．

1. はじめに

日本語には同じ字面でも違う読みをする単語が存在する．このような単語を同形異音語という．例として「辛い」は「カライ」だけでなく「ツライ」と読むこともできる．日本語話者であれば文脈から読み分けをすることは容易だが, 日本語を母語としない者やコンピュータにとっての難易度は高い．

これまでに小林ら [1] は SVM(Support Vector Machine) を用いて同形異音語の読み推定を行ってきた．その際にはデータセットとして現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Written Japanese: BCCWJ)[2], 素性として one-hot ベクトル, nwjc2vec[3], BERT (Bidirectional Encoder Representations from Transformers)[4] による分散表現を用いた．また, この論文では, 分類対象の同形異音語を絞って実験を行っている．具体的には, BCCWJ から 71 個の同形異音語を抽出し, それらの読み推定を行った．

本稿では, 事前学習モデル BERT を用いて, 全単語を対象に読み推定を行う．既存研究との違いは, BERT の fine-tuning を利用した点と, 全単語を対象にした点, 疑似データを利用した点である．

2. 関連研究

2.1 同形異音語の読み推定の関連研究

全単語を対象とした読み推定の研究は, 著者らが知る限り存在しない．対象単語を絞って同形異音語の読み推定を行った研究には, 1 節でも述べた小林らのものがある．この論文では BCCWJ 中に存在する「辛い」(「カライ」「ツライ」)「市場」(「シジョウ」「イチバ」)「生物」(「セイブツ」「ナマモノ」) などの 71 個の同形異音語の読みを, さまざまな素性を用いて推定している．結果はマクロ平均 90.69%, マイクロ平均 96.01%が最高となった．

小林らの研究や本研究で使用された BERT は, 2018 年に Google の Jacob Devlin らが発表したモデルである．fine-tuning を用いることで様々なタスクに応用できることが BERT の特徴であり, この特徴を生かして種々の NLP タスクの最高性能を次々に更新した．

2.2 全単語対象の関連研究

今回, 我々はコーパスに含まれる曖昧な読みを持つ全単語を対象に読みの推定を試みた．単語の「読み」を単語の「語義」とみなせば, 本タスクは all-words WSD(Word

[1] 茨城大学大学院理工学研究科情報工学専攻
Major in Computer and Information Sciences,Graduate School of Science and Engineering, Ibaraki University
[2] 東京農工大学大学院工学研究院先端情報科学部門
Division of Advanced Information Technology & Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology
[3] 茨城大学大学院理工学研究科情報科学領域
Graduate School of Science and Engineering, Department of Computer and Information Sciences, Ibaraki University
[a)] 21nm724l@vc.ibaraki.ac.jp
[b)] kkomiya@go.tuat.ac.jp
[c)] hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

Sense Disambiguation: 語義曖昧性解消) と同形の問題であり, そこでの手法が利用できる. all-words WSD とは, 文書中の全単語を対象に語義ラベルを与えるタスクである. 新納ら [5] はテキスト解析ツール KyTea[*1]を用いてこの問題が解決できることを示した. KyTea は分割されたテキストデータを用いて単語分割を訓練するモデルである. 新納らは訓練データに語義データを加えて学習させることで, 語義曖昧性解消のモデルを構築した. このように, 曖昧性のある全単語にラベルを付与して学習させることで, 全単語を対象とした曖昧性解消システムを構築できる. また, 鈴木ら [6] は概念辞書を用いることで多義語の周辺単語の分散表現を作成し, それらのユークリッド距離を計算することで多義語の語義を推測している. これは, 単語の語義は文脈に依存するという仮定のもとに行われた実験である. その他に, BERT を用いた英語の all-words WSD を行った研究には Jiaju Du et al.[7] のものがある. 彼らは英語の all-words WSD タスクに BERT が有効であることを示している. 具体的には, BERT を用いることで当時の最高性能を 5.2 ポイント上回る結果を残している.

2.3 疑似データを用いることの関連研究

疑似データを用いた研究には, 清野ら [8] や斎藤ら [9] の研究がある. 清野らは疑似データの生成方法や疑似データの生成元について検討している. その結果, CoNLL-2014 において当時の最高性能を記録した. 斎藤らは, 自動生成された疑似正解データを用いて事前学習を行った後, 少量の人手正解データを用いて再度学習させるという手法を試みた. その結果, 文字列の正規化にこの手法が有効であることを示した. また, Xiaojie Wang et al.[10] は, 疑似訓練データと語義タグ付きデータとを組み合わせることが中国語の語義曖昧性解消に効果的であることを示した.

3. 疑似データを用いた同形異音語の読み推定

本研究では, コーパス中の全単語を対象とした読み推定システムを作成する. 読み推定のための教師データとなるような読み情報が正確に付与された日本語コーパスには, 日本語話し言葉コーパス (Corpus of Spontaneous Japanese: CSJ)[11] がある. CSJ は話し言葉をベースに作られたコーパスであり, 音声データを書き起こしているため, 正確な読み情報を得ることが可能である. しかし, CSJ のような音声情報を書き起こすコーパスは構築コストが高く, 大量に用意することは困難である.

そこで本研究では, システムによって自動的に読み情報を付与された疑似データを利用する. 大量の疑似データを用いてモデルを構築したのち, 少量の正解の読み情報が付与されたデータを用いて追加学習を行う手法と, 正解データのみを用いて構築したモデルとの精度を比較した. これにより疑似データの有効性を調査した. その際, 正解データをどの程度利用すれば, 正解データのみを用いたモデルに迫れるかを検証した.

4. データ

本実験では テストデータおよび正解の読み情報が付与された訓練事例として CSJ を利用し, 疑似データの訓練事例として BCCWJ を利用した. BCCWJ は, 書籍全般, 雑誌全般, 新聞, 白書, ブログ, ネット掲示板, 教科書, 法律などのジャンルにまたがって 1 億 430 万語のデータを格納しており, 2022 年 7 月時点で, 日本語について入手可能な唯一の均衡コーパスである. BCCWJ では, 形態論情報をほとんど自動で付与しているが, その一部には人手で解析精度を高めたコアデータが含まれている[*2]. 本研究の実験には非コアデータを利用した. また, BCCWJ は書き言葉であるため, 例えば「日本」が「ニホン」なのか「ニッポン」なのかなど, 正確な読み情報は分からない場合がある. CSJ は, 日本語の自発音声を大量に集め, 品詞等の様々な形態論情報を付加した話し言葉研究用のデータベースである. 音声データの書き起こしによるコーパスであるため, 正確な読み情報が付与されていると考えられる. 上記 2 種類のデータセットに含まれる単語の情報は表 1 のとおりである.

本研究の実験ではテストデータに CSJ を利用し, 訓練に用いた疑似データには BCCWJ を利用した. そのため, 疑似データとテストデータのドメインは異なっている点に注意されたい.

5. モデル

本研究の実験には BERT の fine-tuning を利用した. その様子を図 1 の模式図に示す. 入力 (図 1 の「[CLS] tok1 tok2・・・tokN [SEP]」) はコーパスから整形して抽出したトークン列（6.1 節参照）であり, BERT の 12 の層を経て 768 次元のベクトルへと変換される. このベクトルを識別層 W に入力することで読みラベル R を出力する. ただし, この出力は 15,291(=読みの辞書のサイズ) 次元のベクトルである. このとき, ベクトル中の a 番目の要素は, その単語の読みがラベル a である確率を表している. この確率を参照し, 最も値の大きなラベルをモデルの推定結果とした.

疑似データの効果を調べるため, 訓練データの異なる 5 つのモデルを用意した. 表 2 に各モデルの説明を示す. 全てのモデルは CSJ を分割したテストデータ (6.2 節参照) を用いて評価している.

[*1] http://www.phontron.com/kytea/index-ja.html

[*2] https://clrd.ninjal.ac.jp/bccwj/doc/manual/BCCWJ_Manual_02.pdf

表 1　コーパスの統計情報

	単語の種類数	単語数	曖昧性のある単語の種類数	曖昧性のある単語の出現数
BCCWJ	422,793	123,848,121	4,833	20,081,893
CSJ	62,593	7,142,610	4,551	839,494
全体	442,698	130,990,731	8,950	20,921,387

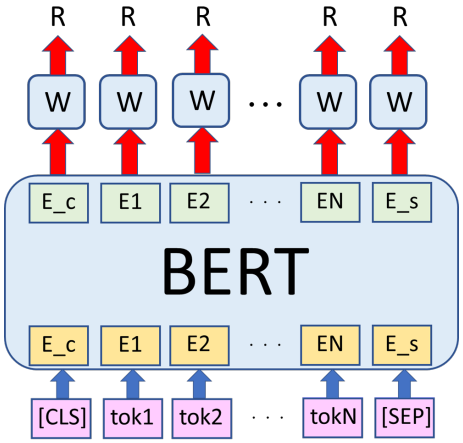

図 1　モデルの模式図

6. 実験

6.1 データの整形

本節では, コーパス中の単語データの整形方法を示す.

まず, 読みに曖昧性のある単語をコーパスから抽出する. これにより読みラベルを定義でき, 読みの辞書を作成することもできる. 実験で用いた全データの読みに関する情報は表 3 の通りである.

次にコーパス中のデータを 1 文毎に区切る. これは, 文毎に BERT へ入力し, 実験を行うためである. ただし, コーパスには単語毎にデータが格納されているため, 各コーパスにおける文の区切り文字を以下のように定義した.

- BCCWJ:「。」「!」「?」
- CSJ:「です」「ます」「た」

このようにして定義された 1 文に様々な情報を付与することで, 文情報 s を作成した. s は以下に列挙する情報から構成されている.

(A) 文中の単語情報 w
(B) BERT への入力となる ID リスト
(C) 読みに曖昧性のある単語の読みラベルリスト
(D) 読みに曖昧性のある単語の読み候補リスト
(E) 曖昧な読みのある単語の位置情報

ここで, 単語情報 w とは, 単語毎に表層や読みなどの情報を与えたものである.

文情報 s の例を, 例文「この本は多くの生物が載っている」を用いて図 2 に示す. なお, 図 2 の (B) には, [CLS] を表す 2 と [SEP] を表す 3 が挿入されている. 本実験では, 読みに曖昧性のある単語についてのみ学習及び推論を行っている. つまり, 例文においては 2 単語目の「本」(「モト」, 「ホン」, 「ボン」, 「ポン」) と 6 単語目の「生物」(「セイブツ」, 「ナマモノ」) を対象に学習・推論を行う. 推論の際には, 読み候補 (図 2 の (D)) の中から最も確率の高いものを選定する.

6.2 実験設定

モデルには東北大から公開されている訓練済み日本語 BERT モデル[*3]を使用した. 使用したモデルのハイパーパラメータのうち, 変更したものは以下の通りである.

- 最適化関数：SGD
- 学習率：10^{-4}
- ミニバッチ数：1

また, CSJ はデータ全体を (訓練データ):(検証データ):(テストデータ)=1:1:8 に分割して利用した.

7. 結果

追加学習に使用するデータ量毎の全単語を対象とした読み推定の正解率は表 3 の通りである.

まず, modelC の読み推定の正解率は 97.97%であり, modelB の正解率が 94.22%であることから, 疑似データのみを学習に利用した読み推定システムは, ターゲットデータである CSJ を学習に利用した場合と比較すると 3.7 ポイントの正解率の差があることが分かる. 学習データとして, 疑似データに加えて 5%の CSJ を追加すると (modelB-C5), 読み推定の正解率は 96.95%となる. さらに追加データを 10%に増やすと (modelB-C10), 正解率は 97.22%となる. 追加データを 20%にした場合 (modelB-C20) の読み推定の正

[*3] cl-tohoku/bert-base-japanese

表 2 モデルの詳細

モデル名	説明
modelC	(C) で学習したモデル
modelB	(B) で学習したモデル
modelB-C5	(B) で学習した後 (C) の 5%で追加学習したモデル
modelB-C10	(B) で学習した後 (C) の 10%で追加学習したモデル
modelB-C20	(B) で学習した後 (C) の 20%で追加学習したモデル

表 3 BCCWJ と CSJ に含まれる読みが曖昧な単語の読み情報

読みが曖昧な単語における読みの種類数	読みが曖昧な単語における読みの総数	読みの候補の平均数
15,291	20,574	2.30

表 4 追加学習に使用する疑似データの量ごとの全単語を対象とした読み推定の正解率

モデル名	正解率 (%)
modelC	97.97
modelB	94.22
modelB-C5	96.95
modelB-C10	97.22
modelB-C20	97.53

解率は 97.53%であった．これらの結果から, 追加データを 20%にしても, 読み推定の正解率は modelC には及ばないことが分かる．しかし一方で, その差はわずかである．CSJ の訓練データ 10%で追加学習を行った modelB-C10 では, modelC との差は 0.75 ポイントに迫っている．反対に modelB-C10 は modelB と 3 ポイントの差があることから, 書き言葉の疑似データで学習させたモデルに 10%の話し言葉データを追加学習させることで, 3 ポイントの正解率の上昇となっていることが見て取れる．つまり, 10%の話し言葉データを追加したことによる話し言葉への領域適応の効果は高く, modelC とほとんど同程度の正解率が達成できることが確認できる．これらの実験から, 書き言葉のコーパスにある自動的に読みを付与したデータを疑似データとして利用すると, 音声書き起こしのコーパスの訓練事例の量を 10%に減らしても, 音声書き起こしのコーパスをすべて利用する場合に比べてほとんど遜色ない読み推定の正解率が得られることが分かった．

8. 考察

modelB に多く見られた誤りとして「私」(「ワタシ」「ワタクシ」「シ」) や「他」(「タ」「ホカ」) が挙げられる．「私」については, 本来「ワタシ」または「シ」と読むべきところを全て「ワタクシ」と読んでいた．「他」については, 誤り全体の約 88.7%が「ホカ」と読むべきところを「タ」と読んでしまい, 残りの約 11.3%が「タ」と読むべきところを「ホカ」と読んでしまっていた．「ワタクシ」は「ワタシ」に比べて形式ばった書き言葉らしい表現である．また,「他」を「タ」と読むのも書き言葉らしいと言える．例えば「その他」という熟語は公用文では「ソノタ」読み,「ソノホカ」とは読まない (「ソノホカ」と読ませたい場合には「その他」とは表記せず, 平仮名で「そのほか」と表記する)．このような書き言葉の特徴が, modelB に「ワタシ」を「ワタクシ」と読ませたり, 多くの「ホカ」と読むべき単語を「タ」読ませたりしたのではないかと考えられる．しかし, modelB-C10 では「私」の誤り率が約 74.9%減少し,「他」の誤り率も約 58%減少していた．このことから, 大量の疑似データがあれば, 少量のターゲット領域データを追加学習させるだけで, 領域適応が可能だと言える．

今後は, CSJ のデータを MeCab にかけることで話し言葉の疑似データを作成し, 今回のモデルと比較することを考えている．

9. おわりに

本稿では, BERT の fine-tuning を用いて読み推定を行った．読み推定の対象は, BCCWJ と CSJ に存在する, 読みに曖昧性のある全単語とした．実験では, ドメイン外の大量の疑似データを用いることで, 構築コストの高いターゲット領域のデータを減らせることを確認した．具体的には, 本来使用する量の 10%程度があれば, モデルの精度に遜色ないことが分かった．

今後は, ターゲット領域の疑似データを用いることで, 更に人手データの量を減らせるのかを検証したい．

謝辞 本研究は 2022 年度国立情報学研究所公募型共同研究（22FC04），JSPS 科研費 22K12145 の助成を受けています．

参考文献

[1] 小林汰一郎, 古宮嘉那子 (2021). SVM を用いた BCCWJ における同形異音語の読み推定. 言語処理学会 第 27 回年次大会 405–409

[2] Makawa, Kikuo and Yamazaki, Makoto and Ogiso, Toshinobu and Maruyama, Takehiro and Ogura, Hideki and Kashino, Wakako and Koiso, Hanae and Yamaguchi, Masaya and Tanaka, Makiro and Den, Yasuharu(2014). Balanced corpus of contemporary written Japanese. LREC2014 345–371

[3] 新納 浩幸，浅原 正幸，古宮 嘉那子，佐々木 稔 (2017). nwjc2vec: 国語研日本語ウェブコーパスから構築した単語の分散表現データ, 自然言語処理, Vol. 24, No. 5，pp.

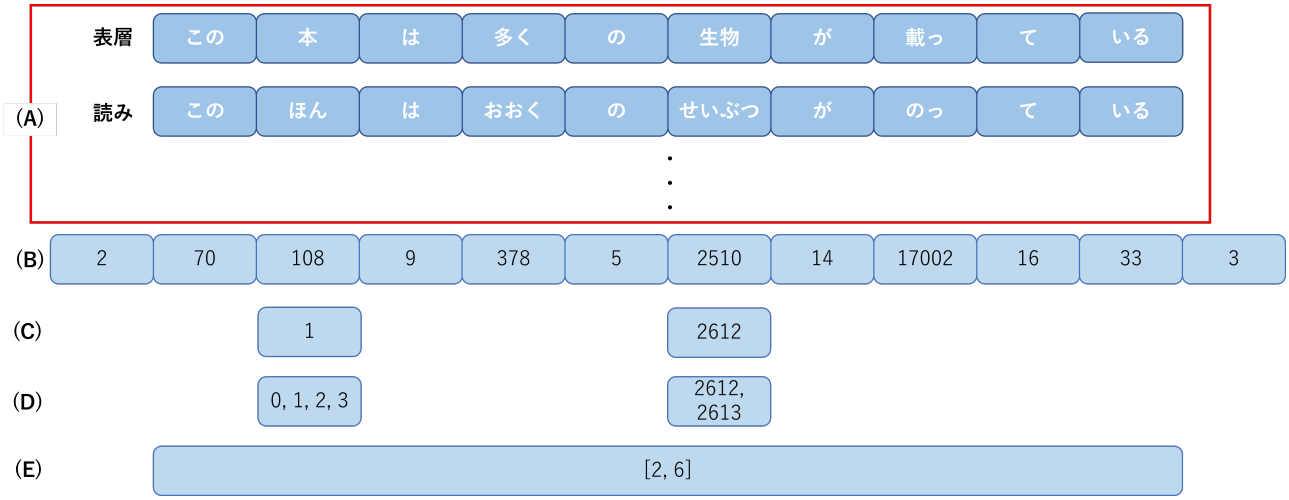

図 2 「この本は多くの生物が載っている」の文情報

705-720, (2017.12).

[4] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina(2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

[5] Shinnou, Hiroyuki and Komiya, Kanako and Sasaki, Minoru and Mori, Shinsuke(2017). Japanese all-words WSD system using the Kyoto Text Analysis ToolKit. Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation. 392–399

[6] 鈴木 類, 古宮 嘉那子, 浅原 正幸, 佐々木 稔, 新納 浩幸 (2019). 概念辞書の類義語と分散表現を利用した教師無し all-words WSD.

[7] Jiadu Du, Fanchao Qi, Maosong San(2019). Using BERT for Word Sense Disambiguation.

[8] 清野 舜, 鈴木 潤, 三田 雅人, 水本 智也, 乾 健太郎 (2020). 大規模疑似データを用いた高性能文法誤り訂正モデルの構築. 言語処理学会 第 26 回年次大会 989–992.

[9] 斉藤 いつみ, 鈴木 潤, 貞光 九月, 西田 京介, 斎藤 邦子, 松尾 義博 (2017). 擬似データの事前学習に基づく encoder-decoder 型日本語崩れ表記正規化. 言語処理学会 第 23 回年次大会 585–588.

[10] Xiaojie Wang, Yuji Matsumoto(2004). Improving Word Sense Disambiguation by Pseudo-samples. IJCNLP 2004: Natural Language Processing – IJCNLP 2004 pp 386–395.

[11] Maekawa, Kikuo(2003). Corpus of Spontaneous Japanese: Its design and evaluation. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition