疑似訓練データを用いた one-shot 設定における同形異音語の読み推定

小林汰一郎 [1]　古宮嘉那子 [2]　新納浩幸 [3]
[1] 茨城大学大学院理工学研究科情報工学専攻
[2] 東京農工大学大学院工学研究院先端情報科学部門
[3] 茨城大学大学院理工学研究科情報科学領域
21nm724l@vc.ibaraki.ac.jp
kkomiya@go.tuat.ac.jp
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

概要

読みに曖昧性を持つ単語を同形異音語と呼ぶ. 本論文で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) に含まれる全ての同形異音語を対象に読み推定を行った. CSJ をテストデータとする場合, 話し言葉データを訓練データとして読み推定のモデルを学習するのが望ましいが, 訓練データ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アノテーションされたドメイン外の大量の疑似データ (BCCWJ のデータ) を用いることで, CSJ の訓練データの量を one-shot まで抑えても精度に大差ないことを示した.

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は困難である.

これまでに小林らは同形異音語の読み推定を行ってきた. 先行研究 [1] では SVM(Support Vector Machine) を用い, データセットとして現代日本語書き言葉均衡コーパス (BCCWJ)[2], 素性として one-hot ベクトル, nwjc2vec[3], BERT [4] による分散表現を用いた. また, 先行研究 [5] では事前学習モデル BERT を用いて, 全単語を対象に読み分類を行った.

本稿では, BERT を用いて全単語を対象に読み推定を行う. 既存研究 [5] との違いは, 自動的にアノテーションされた疑似データを大量に用いることで人手データを one-shot に留め, モデルの構築コストを抑えたことである. このようにして作成されたモデルが, 人手データを大量に使用して学習したモデルの精度に匹敵することを確認した.

2　関連研究

2.1　同形異音語の読み推定の関連研究

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 [1] がある. この論文では BCCWJ 中に存在する 71 単語の読みを, さまざまな素性を用いて分類している. 全単語を対象とした読み推定を行った研究には, 小林ら [5] がある. 本研究では one-shot の設定としたところに違いがある.

2.2　全単語対象の関連研究

本論文で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行った. 全単語を対象にしている点, 読みは意味が違うと異なることが多い点から見ると, 本タスクは all-words Word Sense Disambiguation (WSD, 語義曖昧性解消) と関連がある. 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. WSD では一般に, 単語の意味は文脈に依存すると仮定している. 本研究でもまた, 単語の読みが文脈に依存すると仮定している.

新納ら [6] はテキスト解析ツール KyTea[1)] を用いて all-words WSD 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

1）http://www.phontron.com/kytea/index-ja.html

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語義曖昧性解消システムを構築できる. また, 鈴木ら [7] は概念辞書を用いて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た. また, Jiaju Du et al.[8] は英語の all-words WSD タスクに BERT が有効であることを示した. 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 [9] や斎藤ら [10], Wang ら [11] の研究がある. 清野らは, 文法誤り訂正において, 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 [10] は, スペル訂正タスクにおいて, 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手法が有効であることを示した. また, Wang ら [11] 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. 本研究には, 日本語話し言葉コーパス (CSJ)[12] を利用する. 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学習データとして追加的に利用する. 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正解率を比較することで, 疑似データの有効性を調査した. 本実験では追加学習に CSJ に存在する読みが曖昧な単語を一用例ずつ用い, one-shot の設定で実験を行った.

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データとして CSJ を利用し, 疑似訓練データとして BCCWJ を利用した. 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)]. そのため, 本研究の実験には非コアデータのみを利用した. また, BCCWJ は書き言葉であるため正確な読み情報は分からない場合がある. 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および one-shot の学習データに CSJ を利用し, 疑似訓練データには BCCWJ を利用した. そのため, 疑似訓練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. ただし, 参照する要素は単語によって異なる. 例えば, ある単語の読みがラベル *10*, *11*, *12* に対応するのであれば, 参照する要素は *10*, *11*, *12* 番目のみである.

使用したモデルの説明を以下に示す.

modelC CSJ で学習したモデル
modelB BCCWJ で学習したモデル
modelB-C1s BCCWJ で学習した後 CSJ における one-shot 学習を行ったモデル
modelC1s CSJ における one-shot 学習を行ったモデル

全てのモデルは CSJ を分割したテストデータ (6.2 節参照) を用いて評価した.

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

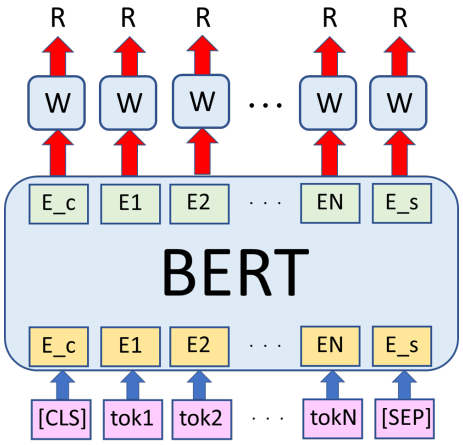

図 1　モデルの模式図

6　実験

6.1　データの整形

本研究では, コーパス中の単語データの整形を以下のように行った. 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し, 読みの辞書を作成する. 表 2 に読みに関する統計情報を示す. 次に, 単語毎に「単語の表記」「BERT における単語の ID」「単語の読み」「単語の読みラベル」「単語の読みの候補リスト」「単語の読みの候補ラベルリスト」「単語の読みの候補数」の 7 つの情報を付与し, 単語情報 w を作成する. さらに, コーパス中のデータを 1 文毎に区切る. 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区切り文字を以下のように定義した.

BCCWJ　「。」「!」「?」
CSJ　「です」「ます」「た」

これらの文毎に「文中の単語情報 w」「BERT への入力となる ID リスト」「曖昧な読みのある単語の位置情報」「読みに曖昧性のある単語の読みラベルリスト」「読みに曖昧性のある単語の読み候補リスト」の 5 つを付与して文情報 s を作成し, これをシステムの入力とした.

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. この訓練データから, 読みに曖昧性のある単語を一用例ずつ抽出し, one-shot 学習用のデータとした. なお, ハイパーパラメータの変更とデータの分割については, 小林ら [5] の研究に倣った.

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を表 3 に示す. まず, modelC1s, modelB, modelC の正解率がそれぞれ 58.40%, 94.22%, 97.97%であることから, 領域外の大量の疑似データを用いた学習による読み推定 (modelB) の正解率は, 対象領域の one-shot 学習によるモデル (modelC1s) より高く, 対象領域の全データを用いた学習によるモデル (modelC) と比べると低いことが確認できる. そ

3)　cl-tohoku/bert-base-japanese

表 2 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 3 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-C1s	97.40
modelC1s	58.40

こで, 学習データとして, 疑似データに加えて CSJ に存在する分類対象の単語一用例ずつを追加すると (modelB-C1s), 読み推定の正解率は 97.40%となった. これは, modelC と比べて 0.57 ポイント下回りこそするが, その差はわずかである. 反対に modelB-C1s 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話し言葉データをごく少量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, 話し言葉データを一用例ずつ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を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データの量を one-shot 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

本章では, modelB と modelB-C1s とを比べて, どのようなデータにおいて改善が見られたのか, その傾向を考察する. 結論としては, modelB-C1s の結果から, 改善されたデータの傾向を読み取ることはできなかった.

傾向を読み取る際, 以下の 3 つを検証した.

- 品詞
- 読みの平均数
- 意味の遠さ

検証にあたって, 正解率の上昇に起因した上位 25 単語 (付録: 表 4) と下位 25 単語 (付録: 表 5) を抽出した. まず, 品詞については, 上位 25 単語のうち 23 単語が名詞, 2 単語が動詞であり, 下位 25 単語では, そのすべてが名詞であった. どちらもほとんど名詞であることから, 両モデルの品詞による傾向の差はないといえる. 次に, 読みの平均数については, 上位 25 単語では 4.12(個/単語), 下位 25 単語では 3.36(個/単語) であった. つまり, 読みの個数には, 上位と下位の間で, 1 単語あたり 0.76 個の差があることがわかる. しかし, 1 単語あたりの読みの個数の差が 1 未満であるため, 大差はないといえる. 次に, 意味の遠さについては, 第一著者の主観で分類した. 例えば「開く」は「ヒラク」や「アク」と読むが, これはほとんど意味が同じである. それに対して「市場」は「シジョウ」や「イチバ」と読むが, これは文脈によって読み分けが必要であるため, 意味が遠いといえる. このように判断した時, 上位 25 単語中意味が遠い単語は「評定」(「ヒョウテイ」「ヒョウジョウ」) のみであり, 下位 25 単語中意味が遠い単語は「市場」(「イチバ」「シジョウ」) と「大勢」(「タイセイ」「オオゼイ」「タイゼイ」), 「出店」(「デミセ」「シュッテン」) の 3 単語である. このような観点で上位と下位を比べても, 精度の差に傾向はみられなかった. これらのことから, modelB-C1s の正解率が modelB を上回った要因は, テストデータと同じ領域の訓練データを追加したという単純な理由によるものと考えられる.

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 (BCCWJ の読みデータ) を用いることで, 構築コストの高い書き起こしによる読みのタグ付きデータ (CSJ の読みデータ) を減らすことができることを確認した. 具体的には, CSJ の読みデータを one-shot まで減らしてもモデルの精度に遜色ないことが分かった.

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．また, 国立国語研究所の異分野融合型共同研究「テキスト読み上げのための読みの曖昧性の分類と読み推定タスクのデータセットの構築」の成果である．

参考文献

[1] 小林汰一郎, 古宮嘉那子. 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, pp. 405–409, 2021.

[2] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language resources and evaluation**, Vol. 48, No. 2, pp. 345–371, 2014.

[3] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. 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. 705–720, 2017.

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

[5] 小林汰一郎, 古宮嘉那子, 新納浩幸ほか. 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. 研究報告自然言語処理 (NL), Vol. 2022, No. 3, pp. 1–5, 2022.

[6] Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. Japanese all-words wsd system using the kyoto text analysis toolkit. In **Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation**, pp. 392–399, 2017.

[7] 鈴木類, 古宮嘉那子, 浅原正幸, 佐々木稔, 新納浩幸. 概念辞書の類義語と分散表現を利用した教師なし all-words wsd. 自然言語処理, Vol. 26, No. 2, pp. 361–379, 2019.

[8] Jiaju Du, Fanchao Qi, and Maosong Sun. Using bert for word sense disambiguation. **arXiv preprint arXiv:1909.08358**, 2019.

[9] 清野舜, 鈴木潤, 三田雅人, 水本智也, 乾健太郎. 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, pp. 989–992, 2020.

[10] 斉藤いつみ, 鈴木潤, 貞光九月, 西田京介, 齋藤邦子, 松尾義博. 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, pp. 585–588, 2017.

[11] Xiaojie Wang and Yuji Matsumoto. Improving word sense disambiguation by pseudo-samples. In **International Conference on Natural Language Processing**, pp. 386–395. Springer, 2004.

[12] Kikuo Maekawa. Corpus of spontaneous japanese: Its design and evaluation. In **ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition**, 2003.

表 4　改善が見られた上位 25 単語

単語	代表的な読み		
私	ワタシ	ワタクシ	シ
後	アト	ノチ	コウ
時	ジ	トキ	ドキ
他	ホカ	タ	アダ
人	ニン	ジン	ヒト
方	カタ	ホウ	ガタ
捉え	トラエ	ツラマエ	
婆	ババ	バア	
京都	キョウト	ミヤコ	
中	ジュウ	チュウ	ウチ
節	セツ	フシ	タカシ
波形	ナミガタ	ハケイ	ナミカタ
形	ガタ	ナリ	カタチ
捉える	トラエル	ツラマエル	
九	ココノ	キュウ	ク
下	シモ	モト	シタ
行ける	イケル	ユケル	
風	フウ	カゼ	プウ
家	ヤ	イエ	カ
行け	イケ	ユケ	
評定	ヒョウテイ	ヒョウジョウ	
車	シャ	クルマ	グルマ
上	ウワ	カミ	ウエ
共	トモ	ドモ	ムタ
拍	ハク	パク	

表 5　改善が見られなかった上位 25 単語

単語	代表的な読み		
日間	ジッカン	カカン	ニッカン
市場	イチバ	シジョウ	
大勢	タイセイ	オオゼイ	タイゼイ
鼻	ビ	ハナ	バナ
湖	ミズウミ	コ	ウミ
姉	アネ	ネエ	
水	ミズ	スイ	ズイ
原	ハラ	ゲン	バラ
味	ミ	アジ	
日	カ	ヒ	ジツ
開く	ヒラク	アク	
縁	ユカリ	エン	ヨスガ
六七	ロクナナ	ロクシチ	
波	ナミ	ハ	パ
梅	バイ	ウメ	
入り	ハイリ	イリ	バイリ
帯	オビ	タイ	
器	キ	ウツワ	
糞	フン	クソ	
幸	ミユキ	サイワイ	サチ
着	ギ	キ	チャク
薬	クスリ	ヤク	グスリ
酒	サケ	シュ	ザケ
肝	カン	キモ	
出店	デミセ	シュッテン	