疑似訓練データによる格助詞の省略に頑健な係り受け解析

石川遼伍　丹羽彩奈　水木栄　岡崎直観
東京工業大学　情報理工学院
{ryogo.ishikawa, ayana.niwa, sakae.mizuki}[at]nlp.c.titech.ac.jp
okazaki[at]c.titech.ac.jp

概要

話し言葉やソーシャルメディア上などで散見される「大学行った」のような格助詞が省略された文は，係り受け解析の精度に悪影響を及ぼす．本稿では，格助詞の一部を人工的に省略した疑似訓練データを用い，係り受け解析精度を向上させるアプローチを二つ紹介する．一つは疑似訓練データで格助詞補完器を学習し，省略された格助詞を補完してから係り受け解析を行うアプローチ，もう一つは疑似訓練データから格助詞の省略を含む文に対応できる係り受け解析器を直接学習するアプローチである．既存の係り受け解析器との比較において，特に後者は格助詞の省略がない文の解析精度を維持したまま，省略文の解析精度を向上させることを確認した．

1　はじめに

近年，スマートフォンの利用率は全年代で一貫して増加しており，LINE や Twitter をはじめとするサービス／アプリの利用者も増加傾向にある [1]．テキストでのコミュニケーションが増えつつあるなかで，ソーシャルメディア等で交わされるメッセージを対象とする研究も多く存在する．そのような研究では，テキスト分析の前処理として係り受け解析を行うことがある [2, 3, 4]．

しかし，ソーシャルメディア上のコミュニケーションでよく見られる「大学行った」（大学（に）行った）のような助詞が省略された文は，係り受け解析精度の低下を招く．池田ら [5] は，助詞が省略されている文に着目し，省略箇所を自動的に推定し，欠落した助詞を補完することにより，係り受け解析の精度を向上させる手法を提案した．しかし，この省略補完は省略された箇所の前後数個の形態素に基づいており，文全体の構造を考慮していない．

日本語の係り受け解析器として，確率モデルに基づく KNP[1] [6] や，Support Vector Machine に基づく Cabocha[2] [7] 等が存在する．また，最近では BERT [8] を利用した手法が提案され，既存の解析器を上回る精度を達成している [9]．一般に，機械学習による手法では訓練データとして係り受け構造が注釈付けされたコーパスが必要である．ゆえに，格助詞が省略されている文（以下格助詞省略文と呼ぶ）の解析精度を向上させるには，格助詞省略文に注釈付けしたコーパスが必要である．ところが，そのようなデータを人手で構築した例は少なく，大規模なコーパスである現代日本語書き言葉均衡コーパス（BCCWJ）[10] にも，さほど収録されていない．

本稿では，人工的に格助詞を省略させた文を疑似訓練データとして利用することにより，訓練データ不足の問題を回避し，格助詞省略文の係り受け解析精度を向上させる手法を二つ提案する．一つは，疑似データで格助詞補完器を，既存のコーパスで係り受け解析器をそれぞれ学習したうえで，推論時には格助詞を補完してから係り受け解析を行うパイプラインの手法（以下 Pipeline 手法と呼ぶ）である．もう一つは，既存のコーパスの一部を格助詞省略文の疑似データに変換し，係り受け解析器を学習したうえで，推論時には格助詞省略文に対して直接係り受け解析を行うエンドツーエンドの手法（以下 E2E 手法と呼ぶ）である．いずれも BERT を基盤とし，事前学習で獲得された知識と周辺文脈の活用を図る．

実験により，既存の解析器と比較すると，特に E2E 手法では非省略文の解析精度を維持しつつ，省略文の解析精度が向上することを確認した．また，人工的な省略文の生成手法は簡素かつランダムなものであっても，省略文に対して有効であることがわかった．Pipeline 手法は E2E 手法には及ばなかったものの，格助詞補完器の F1 スコアは人工的な省略文で 0.94 に達しており，口語的な文に対するテキス

1） https://nlp.ist.i.kyoto-u.ac.jp/?KNP
2） https://taku910.github.io/cabocha/

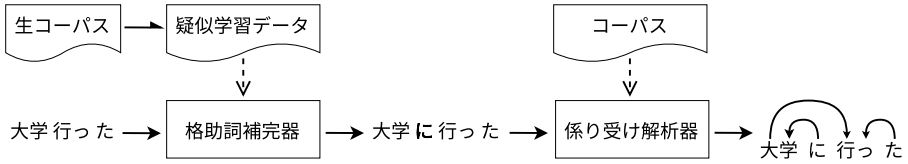


図 1: Pipeline 手法

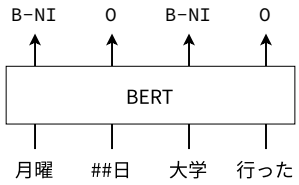


図 2: 格助詞補完モデル

ト正規化の一手法としての応用が期待できる [11].

2 関連研究

船越ら [12] は，音声対話システムにおいて構文解析を困難にする 4 種類の不適格性の解決の一環として，助詞が省略されている文の解析手法を提案した．その手法では，助詞が付属しない名詞からでも動詞に係ること許すように，解析に用いる辞書を拡張した．ただし，評価に用いられた表現や助詞の欠落を含む事例が限られていた．

池田ら [5] は，省略箇所を自動的に推定し，欠落した助詞を補完することによって，係り受け解析の精度を向上させる手法を提案した．助詞の欠落箇所の推定は，任意の品詞間に助詞の欠落があると仮定し，その前後 4 品詞列を特徴量として学習した SVM を用いる．補完は，推定された助詞の欠落箇所の前後最大 4 形態素をクエリとして新聞記事を検索し，助詞別の検索結果の件数をもとにスコアリングして，補完すべき助詞を決定する．このように，池田らの手法では文全体の構文構造が考慮されているわけではない．

3 提案手法

本稿で対象とする格助詞は，船越ら [12] や池田ら [5] を参考に，「が」「を」「に」の 3 種類とした．また，柴田ら [9] と同様に，係り受け解析は単語単位の係り受けを主辞の選択問題として考える．すなわち，サブワードに分割されたトークン列が入力されたとき，各単語の先頭トークンに対して主辞のトークンを予測する問題とする．

3.1 Pipeline 手法

Pipeline 手法では，格助詞補完器によって省略された格助詞の補完を行い，その補完結果に対して係り受け解析を行う（図 1）．

格助詞の補完　格助詞補完器は BERT を疑似訓練データでファインチューニングして構成する．本研究では，格助詞補完を系列ラベリング問題として考える．図 2 に示すように，モデルはサブワードに分割されたトークン列を入力として，あるトークンとその直後のトークンとの間に対象の格助詞のいずれかが挿入されるべきか，何も挿入する必要がないのかを予測する．格助詞補完器は，この予測に基づいて入力の単語列に格助詞を挿入したものを出力する．なお，単語がサブワードに分割されていた場合は，その先頭のトークンのみを考慮する．

疑似訓練データの作成　生コーパスのテキストを形態素解析し，句点で文に分割したのち，品詞と表層形をもとに対象の格助詞をランダムに省略することで疑似データを生成する．省略された格助詞に対応する正解ラベルを付与し，訓練データとする．省略は対象の格助詞ごとに 50%の確率で発生させた．

係り受け解析　通常のコーパスで学習した既存の解析器を用いる．格助詞補完器が出力した単語列を入力にとり，各単語の主辞を予測する．

3.2 E2E 手法

E2E 手法では，既存の係り受け解析モデルを疑似訓練データで学習し，これを用いて省略文の係り受け解析を行う（図 3）．

疑似訓練データの作成　品詞や係り受けの情報が注釈付けされているコーパスをもとに，文ごとに 2 段階でランダムな省略を施し，疑似訓練データを作成する．ある文から格助詞を省略する流れを図 4 に示す．前処理として省略操作に先んじて，省略可能な格助詞を文から抽出する．ここで，省略可能な格助詞とは，他の形態素からの係りがない格助詞を指す．これは，仮に他の形態素からの係りがある格助

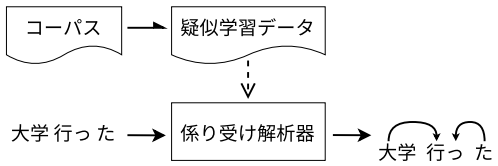


図 3: E2E 手法

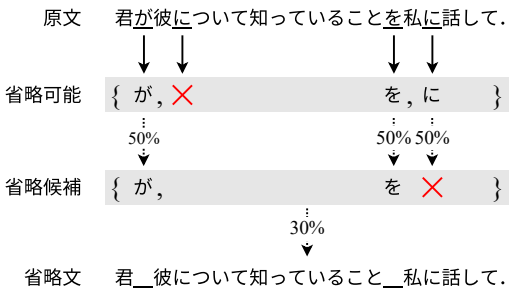


図 4: 人工的な省略文生成の流れ

詞を省略した場合，その文の依存構造木が複数に分割されてしまい，学習時の扱いが煩雑になるためである．本稿で利用した BCCWJ においては，「について」や「に関して」などといった格助詞相当の機能表現の複合語 [13] がこの例として挙げられる．

第 1 段階　省略候補の格助詞を形態素レベルで選択する．省略可能な格助詞それぞれについて，一定の確率で省略候補とする．本稿では，予備実験の結果を踏まえてこの確率を 50%とした．

第 2 段階　実際に省略を適用するのか否かを文レベルで決定する．一定の確率で省略候補を実際に原文から省略し，それによって生じた ID 等のずれを修正する．本稿では，予備実験の結果を踏まえてこの確率を 30%とした．なお，実際に省略が適用された場合，原文はコーパスから取り除かれ，省略文に置き換えられる．

このような 2 段階のプロセスを踏んでいるのは，インフォーマルな文体では格助詞がひとつでも省略されると他の格助詞も省略されやすい一方，フォーマルな文体では全く格助詞が省略されない文も多いという傾向を模擬するためである．$0.5 \times 0.3 = 0.15$ であるから，格助詞ごとに 15%で省略を発生させていることと等価のように思われるが，その場合は 2 段階のプロセスを経るよりも複数の格助詞が省略された文が生成されにくくなる．しかし，単純に格助詞を省略する確率を引き上げてしまうと，訓練データ中の省略文の事例数が多くなりすぎてしまい，解析器は非省略文に対して十分な精度を保つことができなくなってしまう．

4　実験

4.1　データ

係り受け解析の訓練・評価データには，BCCWJ の Universal Dependencies 版（UD Japanese BCCWJ [14]）を用いた．UD Japanese BCCWJ では，あらかじめデータセットが訓練用・開発用・評価用に分割されており，その事例数はそれぞれ 40,801 文，8,427 文，7,881 文で，合計 57,109 文である．本稿では，実際に人間によって書かれた省略文に対する性能を評価するために，元々の分割は用いずに BCCWJ から抽出したすべての格助詞省略文が評価データに含まれるように分割しなおした．この際，各分割の事例数は元々の値をそのまま採用した．

格助詞の省略を明示した注釈付きコーパスが存在しないため，BCCWJ-DepParaPAS [15] の述語項構造を利用して格助詞省略文を抽出した．述語項構造や形態素の情報に基づいて一定の規則を設け，それに従って自動的に大まかな抽出を行ったのち，そこから不適当と判断されるものを人手で取り除いた．自動抽出の基本的な規則は，「ある名詞が述語の項として注釈付けされているにも関わらず，その名詞を含む文節に助詞が存在しない場合は抽出する」というものである．抽出作業の結果，1,745 件の格助詞省略文の事例を得た．評価用の 7,881 文のうち，1,745 文が格助詞省略文で，残りは非省略文である．

4.2　実験設定

Pipeline 手法　生コーパスとしてウィキペディア日本語版の整形済みデータである wiki40b/ja[3] を用いた．形態素解析器には fugashi[unidic-lite][4] を用いた．文ごとに分割した結果，訓練用 13,032,687 件，開発用 716,896 件，評価用 717,627 件の事例を得た．事前学習済みモデルには BCCWJ と品詞体系が一致するように cl-tohoku/bert-base-japanese-v2[5] を使用した．transformers[6] の実装を改変し，BERT モデルを 3 エポックだけファインチューニングした．格助詞補完器のラベリング結果の評価には seqeval[7] を用いた．係り受け解析には後述の Baseline 手法を用いた．

3）https://www.tensorflow.org/datasets/catalog/wiki40b
4）https://github.com/polm/fugashi
5）https://huggingface.co/cl-tohoku/bert-base-japanese-v2
6）https://github.com/huggingface/transformers
7）https://github.com/chakki-works/seqeval

表 1: 係り受け解析の精度（括弧内は有効事例数）

	全体 (7,881)	省略文 (1,745)	非省略文 (6,136)
Baseline	92.78 (7,709)	89.19 (1,695)	93.69 (6,014)
Pipeline	92.83 (7,703)	90.01 (1,691)	93.55 (6,012)
E2E	92.96 (7,709)	**90.42** (1,697)	93.62 (6,012)

表 2: 格助詞補完の精度

	適合率	再現率	F1
が	0.94	0.94	0.94
を	0.96	0.96	0.96
に	0.94	0.93	0.93
全体	0.95	0.94	0.94

E2E 手法　解析モデルには柴田ら [9] による BERTKNP[8] を用いた．事前学習済みモデルには Pipeline 手法と同様に cl-tohoku/bert-base-japanese-v2 を使用した．ファインチューニングは Devlin ら [8] や柴田ら [9] に倣い，3 エポック行った．疑似訓練データの生成はシード値を変えて 5 回行い，評価結果の平均を算出した．

Baseline 手法　ベースラインとして，訓練データ[9] をそのまま用いて解析器 BERTKNP を学習した場合を採用した．事前学習済みモデルやエポック数は E2E 手法と同様とした．

評価指標　係り受け解析の評価には，CoNLL 2017 Shared Task [16] の評価スクリプトを用いた[10]．解析器が出力した依存構造木において，根が存在しない場合や，依存関係にループが存在する場合など，構文木として無効である場合は，規定に従いその事例に対するスコアを 0 として計算した．

4.3　実験結果

実験結果を表 1 に示す．Baseline 手法の省略文に対する精度は非省略文に対するそれよりも 4.5 ポイント低く，格助詞の省略が係り受け解析の精度低下を招くことが改めて示された．Pipeline 手法を見ると，Baseline 手法と比較して省略文に対しては 0.82 ポイントの向上，非省略文に対しては 0.11 ポイントの悪化という結果で，Pipeline 手法では省略文の解析精度を向上させる効果が大きくないことがわかった．E2E 手法では省略文に対する精度が 90.42 ± 0.14 と 3 手法のうちで最も良く，Baseline 手法と比較して 1.2 ポイントの精度向上を達成しており，E2E 手法の省略文に対する有効性を確認することができた．また，E2E 手法の非省略文に対する精度は 93.62 ± 0.15 で，Baseline 手法の精度と大差がなく，E2E 手法が非省略文の解析精度にほとんど悪影響を与えないことも確かめられた．E2E 手法による具体的な係り受け解析の改善事例は付録 A で紹介する．Pipeline 手法と E2E 手法を比較すると，総合的に E2E 手法のほうが有効であった．これは，人工的な省略文をエンドツーエンドで学習させることにより格助詞が省略されていたとしても BERT が文脈を広くとって解析を行えていることや，Pipeline 手法のように格助詞補完器からの誤りの伝播が発生しえないためであることが推測される．

表 2 に Pipeline 手法の格助詞補完器の人工的な省略文における精度を示す．いずれの格助詞についても F1 スコアは 0.93 以上と，人工的な省略文に対しては高い精度で補完ができている．

5　結論

本稿では，口語的な文で頻繁にみられる格助詞の省略された文が係り受け解析の精度に悪影響を及ぼすという問題に対して，省略文に注釈付けしたデータが入手困難な状況において，疑似訓練データを作成し，格助詞補完器や係り受け解析器の学習に用いることで，省略文の解析精度を向上させる手法を提案した．人間によって書かれた格助詞省略文を BCCWJ から抽出して構築した評価データにおいて，E2E 手法は既存の解析器より 1.2 ポイント高い精度を示し，非省略文に対しても既存の解析器と同等の精度を維持できることが確認できた．E2E 手法では，人工的な格助詞の省略を 2 段階のランダムな試行によって行う簡潔な手法を提案し，その有効性を確認した．一方で，Pipeline 手法は十分な効果がみられなかったが，疑似訓練データによる格助詞補完器自体の精度は高く，翻訳タスク等におけるテキスト正規化の手段として活用できる．今後は，より適切な対象格助詞や省略確率の特定，実際にソーシャルメディア上で発信されているテキストでの評価等を行う予定である．

謝辞

本研究は JSPS 科研費 19H01118 の助成を受けたものです．

8）https://github.com/ku-nlp/bertknp
9）再分割後のもの．すべて非省略文からなる．
10）https://github.com/ufal/conll2017/blob/master/evaluation_script/conll17_ud_eval.py

参考文献

[1] 総務省. 令和2年度情報通信メディアの利用時間と情報行動に関する調査, 2021. https://www.soumu.go.jp/iicp/research/results/media_usage-time.html.

[2] Michal Ptaszynski, Fumito Masui, Yuuto Fukushima, Yuuto Oikawa, Hiroshi Hayakawa, Yasunori Miyamori, Kiyoshi Takahashi, and Shunzo Kawajiri. Deep learning for information triage on twitter. **Applied Sciences**, Vol. 11, No. 14, 2021.

[3] 尾崎航成, 向井宏明, 松井くにお. SNSにおける不適切投稿の検知. 情報処理学会 第82回全国大会講演論文集, pp. 621–622, 2020.

[4] 峰松優, 藤淵俊王, 有村秀孝. ソーシャルビッグデータを活用した放射線被ばくに対する不安意見の解析システムの開発. 日本保健物理学会, Vol. 55, No. 1, pp. 15–22, 2020.

[5] 池田和史, 柳原正, 服部元, 松本一則, 小野智弘. 口語文書の解析精度向上のための助詞落ち推定および補完手法の提案. 情報処理学会研究報告, Vol. 2010-DBS-151, No. 39, pp. 1–8, 2010.

[6] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In **Proceedings of the Human Language Technology Conference of the NAACL, Main Conference**, pp. 176–183, 2006.

[7] Taku Kudo and Yuji Matsumoto. Japanese dependency structure analysis based on support vector machines. In **2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora**, pp. 18–25, 2000.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[9] 柴田知秀, 河原大輔, 黒橋禎夫. BERTによる日本語構文解析の精度向上. 言語処理学会 第25回年次大会, pp. 205–208, 2019.

[10] 前川喜久雄. KOTONOHA『現代日本語書き言葉均衡コーパス』の開発 (<特集>資料研究の現在). 日本語の研究, Vol. 4, No. 1, pp. 82–95, 2008.

[11] 笠原要, 斉藤いつみ, 浅野久子, 片山太一, 松尾義博. テキスト正規化技術を用いたcgm日本語テキスト翻訳. 言語処理学会 第21回年次大会, pp. 804–807, 2015.

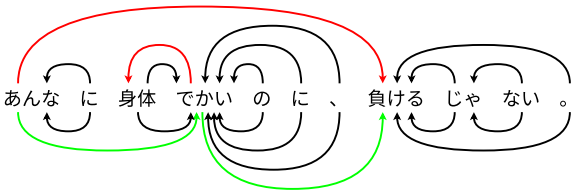
[12] 船越孝太郎, 徳永健伸, 田中穂積. 音声対話用構文解析器の頑健性の評価. 情報処理学会研究報告, Vol. 2002-NL-152, No. 104, pp. 35–41, 2002.

[13] 浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇有吾, 松本裕治. Universal dependencies日本語コーパス. 自然言語処理, Vol. 26, No. 1, pp. 3–36, 2019.

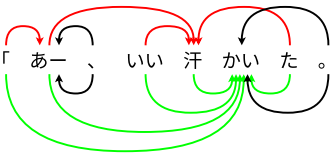
[14] 大村舞, 浅原正幸. 現代日本語書き言葉均衡コーパスのUniversal Dependencies. 言語資源活用ワークショップ発表論文集 = Proceedings of Language Resources Workshop, No. 2, pp. 133–143, 2017.

[15] 浅原正幸, 大村舞. BCCWJ-DepParaPAS:『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化. 言語処理学会 第22回年次大会, pp. 489–492, 2016.

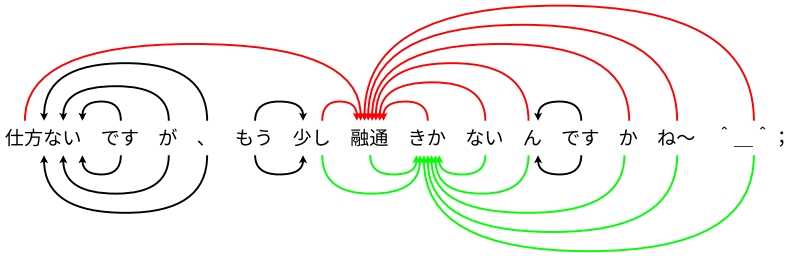
[16] Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In **Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**, pp. 1–19, 2017.



(a) 係り受けがループ



(b) 根が複数



(c) 根を誤って予測

図 5: E2E 手法による改善例

A 係り受け改善例

Baseline 手法では係り受け解析に失敗するものの，E2E 手法では完全に正しく解析できた事例を図 5 に示す．赤矢印が Baseline 手法で解析したときに誤って予測されたもの，緑矢印が E2E 手法により正解を予測できたものを表す．いずれの事例でも，格助詞の省略箇所前後の単語が関係する係り受けの誤りを修正できていることがわかる．特に図 5a・図 5b はベースライン手法では木構造の制約を満たさない出力をしてしまった事例である．