2W-05

漸進的係り受け解析における BERT を用いた未入力文節との構文的関係の同定

橋本 優希 †, a)　　大野 誠寛 †, b)　　松原 茂樹 ‡
東京電機大学未来科学部 †　　名古屋大学情報連携推進本部 ‡

1 はじめに

同時通訳や字幕生成などの音声言語システムでは，入力と同時的に処理することが求められる．このようなシステムにおいて構文的情報を利用するには，音声入力の途中で随時，構文構造を提供できる必要がある．

このような要請に答えるため，文節が入力されるごとに解析を実行し，係り先が入力されていない文節に対して，その係り先は未入力であることを明示した係り受け構造を出力するという漸進的解析手法（以下，大野らの手法）が提案されている [1]．さらに，大野らの手法の出力構造を入力として，係り先が未入力である文節が複数あるときは，それらの係り先が同一か否か（すなわち，未入力文節との構文的関係）を同定する手法（以下，相津らの手法）が提案されている [2]．この手法では，SVM やロジスティック回帰，最大エントロピー法を用いており，精度は必ずしも十分ではない．

本稿では，相津らの手法の精度向上を目指し，同手法において係り先が同一か否かの二値分類を行う際に BERT を用いた手法を提案する．

2 漸進的係り受け解析の出力構造

大野らの手法は，文節が入力されるごとに解析を実行し，係り先が入力されていない文節に対して，その係り先は未入力であることを明示した係り受け構造を出力することを目的としている．図 1 は，文「私は友達が新しい本を買ったのを知っている」の「本を」までが入力された段階で大野らの手法が出力する構造を示しており，「私は」，「友達が」，「本を」の係り先が未だ入力されていないということを示している．これにより，既入力文節内の「新しい本を」が構文的まとまりを構成することがわかる．

一方，係り先が未入力である文節が複数存在したとき，それぞれの文節が異なる未入力文節に係ることもあれば，同一の未入力文節に係ることもある．各文節の係り先が同一か否かを同定できれば，構文的なまとまりをより詳細に捉えることが可能となり，音声言語システムに対して，より早期に，より豊かな構文情報を提供できる。

本研究では，大野らの手法による漸進的係り受け解析の結果を入力として，係り先が未入力である文節が 2 つ以上存在したとき，それらの係り先が同一であるか否かを決定することにより，図 2 のような係り受け構造の同定を試みる．図 2 の文節「友達が」と「本を」は同一の未入力文節（未入力文節 A）に，文節「私は」は「友達が」や「本を」とは異なる文節（未入力文節 B）に係る．このような係り受け構造を同定できれば，「友達が新しい本を」と未入力文節 A からなる文節列が構文的まとまりを構成していることがわかる．

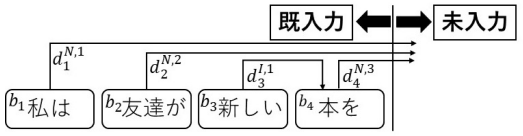

図 1: 大野らの手法が出力する係り受け構造

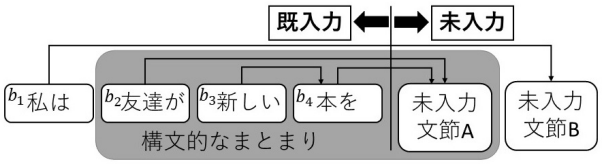

図 2: 本手法により同定される係り受け構造

3 漸進的係り受け解析における未入力文節との構文的関係の同定

本手法の問題設定は相津らの手法 [2] と同じである．すなわち，文節列 $b_1 b_2 \cdots b_m$ からなる文を解析する際，文節 $b_x (1 \leq x \leq m-1)$ が入力されるたびに，それまでに入力された文節列 $B_x = b_1 b_2 \cdots b_x$ と大野らの手法の出力する係り受け構造 D_x とを入力とし，係り先が未入力の係り受け関係が複数ある場合は，それらの係り先が同一か否かを二値分類として出力する．

ここで，大野らの手法が出力する D_x は，文節列 B_x に対する図 1 の形をした係り受け構造であり，係り先が未入力の係り受け関係 $d_k^{N,\alpha} (1 \leq k \leq x, 1 \leq \alpha \leq x)$ と，係り先が既入力の係り受け関係 $d_k^{I,\beta} (1 \leq k \leq x, 1 \leq \beta \leq x-1)$ の集合として定義されるものとする．k は係り元文節の番号を，N は係り先が未入力であることを，I は係り先が既入力であることを意味する．α は係り先が未入力の係り受け関係の中で，また，β は係り先が既入力の係り受け関係の中で，それぞれ係り元文節の番号で昇順に並べた際の順番を示す．例えば，図 1 の係り受け構造 D_4 は，$D_4 = \{d_1^{N,1}, d_2^{N,2}, d_3^{I,1}, d_4^{N,3}\}$ と表記される．

本手法のアルゴリズムを以下に示す．

1. D_x において，係り先が未入力の係り受け関係 $d_k^{N,\alpha}$ の数 L（α の最大値）を集計し，$L = 1$ の場合は終了する．$L \geq 2$ の場合は手順 2 の判定を $\alpha = 1$ から $\alpha = L-1$ まで $L-1$ 回繰り返す．
2. 係り先が未入力の係り受け関係 $d_{k'}^{N,\alpha} (1 \leq k' \leq x-1)$ と，$d_{k''}^{N,\alpha+1} (2 \leq k'' \leq x)$ の両者の係り先が同一であるか否かの二値分類を BERT を用いて行う．

図 1 を例にすると「本を」が入力された段階での大野らの出力構造は $D_4 = \{d_1^{N,1}, d_2^{N,2}, d_3^{I,1}, d_4^{N,3}\}$ となる．このうち，係り先が未入力である係り受け関係は $d_1^{N,1}, d_2^{N,2}, d_4^{N,3}$ の 3 つである（$L = 3$）．まず，$d_1^{N,1}, d_2^{N,2}$ の判定を，次に $d_2^{N,2}, d_4^{N,3}$ の判定を行う．

3.1 BERT を用いた係り先が同一か否かの二値分類

相津らは内元ら [3] の素性のうち，当該 2 文節の語彙情報に関する素性を用いて，係り先が未入力である文節の係り先が同一であるか否かの二値分類（上記手順 2 での判定）を SVM，ロジスティック回帰，最大エントロピー法の 3 つの機械学習により行っている．一

Identification of Syntactic Relations with Non-Inputted Words using BERT in Incremental Dependency Parsing

Yuki Hashimoto†, a), Tomohiro Ohno†, b), Shigeki Matsubara‡
† School of Science and Technology for Future Life, Tokyo Denki University.
‡ Information and Communications, Nagoya University
a) 19fi092@ms.dendai.ac.jp
b) ohno@mail.dendai.ac.jp

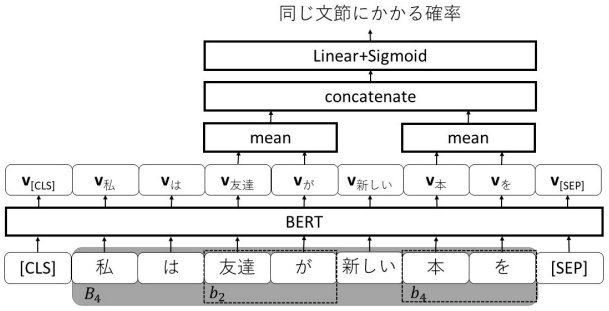

図 3: 本手法における BERT モデル

方，本手法では，文脈情報を考慮するため，当該 2 文節のみならず，入力済みの全文節を使用し，事前学習された BERT を fine-tuning して作成する．

図 3 に本手法における BERT モデルの概要を示す．図 3 では，図 1 の入力に対して，文節 b_2「友達が」と b_4「本を」の係り先が同一か否かの二値分類を行っている．入力文節列 B_x（図 3 では $x = 4$）の先頭に [CLS]，末尾に [SEP] を付与し，サブワード分割したものを BERT に入力する．判定する 2 文節をそれぞれ b_i, b_j（図 3 では $i = 2, j = 4$）とすると，文節 b_i をサブワード分割したトークン列 $t_1^i t_2^i \cdots t_p^i (p \geq 1)$ に対応する BERT の出力 $\mathbf{v}_1^i, \mathbf{v}_2^i, \cdots, \mathbf{v}_p^i$ の平均 $\mathbf{v}_i$ を求め，文節 b_j からも同様に $\mathbf{v}_j$ を求める．次に $\mathbf{v}_i$ と $\mathbf{v}_j$ を連結し，1 層の Linear 層と Sigmoid を介して，文節 b_i と b_j が同じ文節に係る確率を出力する．その確率が 0.5 以上で「係り先は同一」，0.5 未満で「係り先は同一でない」と判定する．

4 評価実験

本手法の有効性を確認するために，日本語講演データを用いて評価実験を行った．

4.1 実験概要

実験データとして，同時通訳データベース [4] に収録されている日本語講演音声の書き起こしデータ（形態素情報，文節境界情報，節境界情報，係り受け情報付）を使用した．なお，係り受け情報をテスト時に使う際は大野らの手法の出力結果に置き換えて使用した．実験は全 16 講演を用いた交差検定により実施した．すなわち，1 講演をテストデータとし，残りの 15 講演を学習データとする実験を 16 回繰り返した．ただし評価では，大野ら [1] における評価用データと同じ 14 講演 (1,714 文，20,707 文節) を使用した．

評価には，係り先が未入力の文節に対する係り先の同定性能として再現率，適合率を用いた．ここで，本手法は係り先が未入力である文節について，その係り先文節を具体的に決めるわけではないため，正解と出力結果の係り先が一致するかを単純には判定できない．相津ら [2] と同様に，本手法の出力から擬似的な係り先文節を用意し，一致する係り受け関係の数が最も多くなるように正解と出力の係り先文節を動的計画法を用いて対応付け，その結果をもとに，正解と出力結果の係り先が一致するか否かを判定した．

比較手法として，相津らの手法のうち、ロジスティック回帰を用いたものを再実装し用意した．本手法のモデルは PyTorch[*1] を用いて実装し，BERT の事前学習モデルは東北大学が公開しているモデル[*2] を用いた．ロジスティック回帰は scikit-learn[*3] を用いた．

[*1] https://pytorch.org/
[*2] https://github.com/cl-tohoku/bert-japanese
[*3] https://scikit-learn.org/stable/

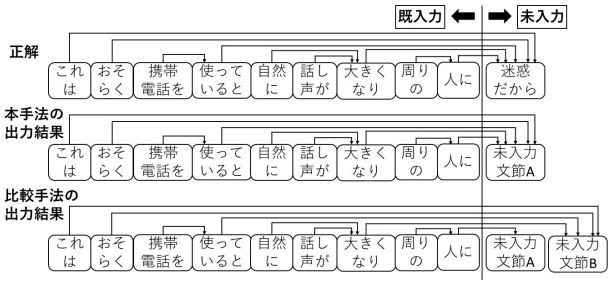

図 4: 本手法の成功例

表 1: 実験結果

手法	再現率	適合率	F 値
相津らの手法	67.48%	70.06%	68.75
(ロジスティック回帰)	(34,615/51,299)	(34,615/49,406)	
本手法	69.09%	71.74%	70.39
(BERT)	(35,442/51,299)	(35,442/49,406)	

4.2 実験結果

表 1 に実験結果を示す．再現率，適合率及び F 値において，本手法は比較手法を約 1.6 ポイント上回っており，本手法の有効性を確認した．

図 4 に，未入力文節との構文的関係の同定において，比較手法が失敗し，本手法が成功した例を示す．比較手法では,「大きくなり」と「人に」の 2 文節が同じ文節に係るか否かの二値分類に失敗している．本手法では,「大きくなり」と「人に」以外の文節も BERT に入力しており，文脈を加味できると考えられるため，この二値分類に成功したものと考察できる．

一方，本手法が失敗した例としては，大野らの手法の出力構造（すなわち，本手法の入力）が正しくないことが原因と考えられる例が多く見られた．そこで，本手法の同定性能を単独で評価するため，日本語講演データに付与されている正解の係り受け構造から，図 1 の構造を抽出し，本手法への入力とした場合の実験を実施した．正解の構造を入力としているため，再現率と適合率は一致することになり，実験の結果，比較手法は再現率と適合率はともに 80.51% (41,303/51,299) となり，本手法は 82.68% (42,414/51,299) となった．正解の係り受け構造を入力に用いると，比較手法との同定性能の差はさらに広がっており，入力構造の正確さが向上すれば，本手法がより有効に働くことを確認した．

5 おわりに

本論文では，未入力文節との構文的関係を BERT を用いて同定する手法を提案した．実験の結果，BERT を用いることの有効性を確認した．今後は，入力に用いる係り受け構造の正確さを改善することで更なる精度向上を図りたい．

謝辞 本研究は，一部，科学研究費補助金基盤研究 (C) No. 19K12127 により実施した．

参考文献

[1] 大野ら，“文節間の依存・非依存を同定する漸進的係り受け解析,” 信学論, J98-D(4), pp. 709–718, 2015.
[2] 相津ら，“漸進的係り受け解析における未入力文節との構文的関係の同定,” 情報処理学会第 82 回全国大会講演論文集, No. 2, pp. 441–442, 2020.
[3] 内元ら，“最大エントロピー法に基づくモデルを用いた日本語係り受け解析,” 情処学論, 40(9), pp. 3397–3407, 1999.
[4] S. Matsubara et al.,“Bilingual Spoken Monologue Corpus for Simultaneous Machine Interpretation Research” Proc. LREC2002，pp. 154–159，2002.