名詞役割入れ替えに頑健な日本語含意関係認識

雨宮正弥[1] 増川哲太[1] 仲田明良[1] 狩野芳伸[1]
[1]静岡大学情報学部
{mamemiya, tmasukawa, anakada, kano}@kanolab.net

概要

含意関係認識に必要な要素は、語彙、構文、推論など様々であるが、たとえば語彙が共通しているが主語と目的語が入れ替わり文意が変わって矛盾となるべきような場合、現状の深層学習モデルでは分類性能が著しく低い。本研究では、非矛盾の文章ペア中の名詞を入れ替えることで矛盾とみなせる文章データを自動作成し、既存の含意関係認識データセットと合わせて BERT をファインチューニングした。学習と評価には、既存の含意関係認識データセットに加え、Wikipedia と新聞記事から抽出したものを組み合わせて用いた。これにより、従来から高い性能が報告されているタイプの含意関係認識の性能は保持しつつ、日本語の文内で名詞を入れ替えたときに意味が変わり矛盾となったかを判別できることを示した。

1 はじめに

含意関係認識は、前提文と仮定文の文ペアについて、前提文が正しいとしたときに仮定文が正しいと言えれば含意、間違えていると言えれば矛盾、どちらとも言えない場合は中立と判定するタスクである。

日本語では日本語 SNLI データセット（JSNLI）[1], 本語構成的推論・類似度データセット（JSICK）[2], Japanese Realistic Textual Entailment Corpus （JRTEC）[3]などの含意関係認識データセットが作成され、事前学習済みの大規模深層言語モデルをファインチューニングすることで 9 割近い分類性能が報告されている[1][2][3]。一方で、既存の深層言語モデルによる日本語の含意関係認識では語順の変化に対応できていないと報告されている[2]。含意関係認識は実際には語彙､構文､推論などその認識に必要な要素が様々であり、それらをまとめた全体性能のみでは性能を判断しがたい。特に、語彙は共通のまま主語と目的語を入れ替えたような場合、現状の深層学習モデルでは分類性能が著しく低い。こうした入れ替え前後の文意の違いを捉えることは非常に重要であり、実現すれば深層学習モデルの根本的な処理能力の向上が期待できる。

本研究では、このような文章内で名詞を入れ替えたときに文意が変わる場合に着目し、自動で入れ替えたデータを生成しファインチューニングに用いる。名詞入れ替えデータの生成においては、入れ替え後の文の不自然さなどほかの手掛かりで判別できないように、品詞細分類や格フレームを用いたいくつかのフィルタを適用した。

名詞入れ替えデータと既存の含意関係認識データとを合わせて学習することで、これまでの含意関係認識の性能を保持しつつ、名詞を入れ替え文意が変わる文にも対応できることを示した。訓練と評価は、各種の既存含意関係認識データセット、Wikipedia、新聞記事データセットを組み合わせて実行し、データセットによるバイアスがないか確認した。

2 関連研究

日本語の含意関係認識データセットには、JSNLI [1]、JSICK[2]、JRTEC[3]、RITE[4]、RITE-2[5]、RITE-Val[6]、Textual Entailment 評価データ[7]などがある。

JSNLI[1]は、英語の大規模含意関係認識データセットである SNLI[8]を日本語に翻訳したもので、事前学習済み BERT [9]をファインチューニングし、評価値 0.929 の分類性能を達成した[1]。

JSICK[2]は、多様な言語現象を含む英語の含意関係認識データセットである SICK[10]を日本語に翻訳したもので、事前学習済み BERT をファインチューニングすることで Accuracy 84.0 の正答率を達成した[2]。また、項（名詞句）の語順を入れ替えても意味内容が変化しない場合に入れ替えをすると、10%程度正答率が下がることを示した。

JRTEC[3]は、根拠付きアノテーションをした含意関係認識データセットで、BERT のファインチューニングにより F 値 92.4 の分類性能を達成している[3]。

RITE[4]、RITE2[5]、RITE-VAL[6]は、NTCIR の shared task で使用されたデータである。RITE-VAL において Ishii らは Macro F1 77.96 を達成している[11]。

Textual Entailment 評価データ[7]は、クラスが◎、○、△、×の 4 つから構成される含意関係認識データセットである。また、推論の要因として包含、語彙（体言）、語彙（用言）、構文、推論の 5 つで各サンプルが分類されている。

3 提案手法

含意関係認識と一口にいっても、実際にはさまざまな要素がある。たとえば Textual Entailment 評価データでは、包含、語彙（体言）、語彙（用言）、構文、推論の 5 つに大きく分類している。

表 1 に Textual Entailment 評価データの含意関係認識の例を示す。なお、Textual Entailment 評価データは◎、○、△、×の 4 つのクラスから構成されるが、他の含意関係認識データセットに合わせ、本研究では◎を含意、○と△を中立、×を矛盾とする 3 ラベルとして扱う。

表 1　Textual Entailment 評価データの例

要素	ラベル	文章
語彙（体言）	含意	前提：肉まんを食べた。 仮説：中華まんを食べた。
構文	矛盾	前提：カマキリは昆虫だ。 仮説：昆虫はカマキリだ。
推論	中立	前提：疲れていませんか。 仮説：休憩しましょう。

本研究の目的は、表 1 内の「構文」例に当たる、日本語の文内で名詞を入れ替えたときに、文意が変わったかどうかを深層言語モデルに判別させることである。原文と名詞を入れ替えた文とのペアを用意し、原文を前提、入れ替え文を仮定とし、正解ラベルを「矛盾」とする含意関係認識タスクとして解く。

文内の任意の名詞ペアを入れ替えてしまうと、日本語として不自然、あるいは単語の並びの頻度が低いかどうかなどで容易に判別できてしまう可能性がある。そこで品詞細分類が同じ名詞のみを入れ替え対象とし、以下で詳述するように、述語とそれが格関係をもつ語(項)を記述した京都大学格フレーム[11]を用いて、入れ替え後もできるだけ日本語として成立しうる文になるようにした。

この名詞入れ替えデータと既存の含意関係認識データセットとを組み合わせて学習・評価を行った。

3.1 名詞入れ替えデータの作成

原文となる日本語文章データとして、データ量が多い日本語 Wikipedia データ[i]と中日新聞の過去 30 年分の記事データから一部を抽出して使用した。

Wikipedia、新聞記事の各文に対し、以下の 1 から 3 の操作を行い、名詞入れ替えデータを作成した。40 文字以下の文のみを対象とし、名詞の入れ替えは 1 文につき一度とした。

1. JUMAN++(pyknp 0.6.1)[ii]で形態素解析をし、文中の名詞を特定する。名詞が連続している場合は 1 つの名詞として扱う。
2. 1 で得られた名詞のうち、品詞細分類が同じ最初の二つを入れ替えた文を作る。
3. 2 で得られた文を日本語構文解析器 KNP(pyknp 0.6.1)[ii] [12]で解析し述語と対応する格を取得した。この結果を京都大学格フレーム Ver 2.0[iii] [11]と照らし合わせ、全ての述語と格の組み合わせがデータに存在する場合のみを採用し、他は除外した。

3.の操作により、入れ替え後の文が日本語として不自然なものが除かれると期待する。例えば原文「鉛筆で字を書く。」に対し「鉛筆」と「字」を入れ替えると「字で鉛筆を書く。」と不自然な内容になるが、これを除外することができた。

表 2 に、入れ替え後の結果を Wikipedia、新聞記事それぞれから 100 件サンプリングして目視で検査した結果を示す。

表 2　名詞入れ替えデータの人手分類の統計

データ	Wikipedia	新聞記事
元文章数	2924560	241117
入れ替え後文章数	33884	10599
意味が変わる	81%	84%
意味が変わらない	12%	11%
不自然な文	7%	5%

作成した名詞入れ替えデータのうち、入れ替え後も文意が変わらないサンプル、文意が変わるサンプル、入れ替え後は日本語として不自然なサンプル、の三

[i] https://dumps.wikimedia.org/jawiki/latest/jawiki-latest-pages-articles.xml.bz2
[ii] https://pyknp.readthedocs.io/en/latest/
[iii] https://www.gsk.or.jp/catalog/gsk2018-b/

種に分類した。ここでいう不自然な文には、構文的に不自然な場合も含まれている。

表 3 に入れ替え後の事例を挙げる。文意が変わる例では、「天気」と「季節」が前提と仮説の文で入れ替わっており、違う文意になっている。文意が変わらない例では「追加」と「店舗閉鎖」が入れ替えられているが、「追加の店舗閉鎖」と「店舗閉鎖の追加」は同じ意味なので入れ替え後の文意が変化していない。不自然な日本語の例では、2 つの名詞の間に助詞「の」が入っているため「お別れの会」全体で一つの名詞と認識されず、「後日お別れ」と「会」が入れ替わる不自然な結果となった。

表 3　入れ替え文章の例

ラベル	文章
文意が変わる	前提：天気が変わりやすい季節。
	仮説：季節が変わりやすい天気。
文意が変わらない	前提：追加の店舗閉鎖は行わない。
	仮説：店舗閉鎖の追加は行わない。
不自然	前提：後日お別れの会を開く予定。
	仮説：会の後日お別れを開く予定。

3.2 含意関係認識器

含意関係認識データの先行研究では、含意関係認識器として BERT が最も多く使用されていたため、本研究でも比較のため BERT を用いる。本研究では BERT の事前学習済みモデルをファインチューニングして含意関係認識器を構築する。入力は「[CLS]仮説[SEP]前提[SEP]」の形式で、[CLS]に対応する出力を用いて含意、中立、矛盾の 3 クラス分類を行う。

4 実験

既存の含意関係認識データだけでファインチューニングした BERT と、既存の含意関係認識データに加えて名詞入れ替えデータで学習した BERT を用いて、含意関係分類タスクを行い評価した。

4.1 データセット

JSNLI と JSICK の 2 つの含意関係認識データセットと、3.1 の手順で作成した名詞入れ替えデータ（**Wikipedia shuffle** および**新聞記事 shuffle**）を、BERT ファインチューニングに使う学習データ・テストデータとして使用した。表 4 に各データセットの統計を示す。JSNLI と JSICK は訓練データとテストデータに分けて提供されているため、それぞれのデータ数を記載した。

表 4　データセットの統計（文ペア数）

データ	矛盾	含意	中立
JSNLI train	178700	176309	177996
JSNLI test	1156	1432	1328
JSICK train	823	1091	3086
JSICK test	797	1088	3042
Wikipedia shuffle	33884	-	-
新聞記事 shuffle	10599	-	-

4.2 実験設定

以下の 4 通りの学習データセットで BERT をファインチューニングした。名詞入れ替えデータを学習する際は、含意のデータ数と矛盾のデータ数が同じくらいになるよう調整し、ランダムサンプリングした。

- JSNLI データセットの学習データ
- JSICK データセットの学習データ
- JSNLI データセットの学習データのうち矛盾ラベルからランダムに文ペア 2 万件を削除したもの（含意約 17 万、矛盾約 15 万、中立約 17 万）+ Wikipedia 入れ替えデータ（矛盾約 2 万）
- JSICK データセット（含意約 1100、矛盾約 800、中立約 3000）+ Wikipedia 入れ替えデータ（矛盾約 300）

学習データは学習用・検証用で 4:1 に分割して学習した。学習や推論には Hugging Face の BertForSequenceClassification[iv]を、BERT の事前学習モデルは東北大学の乾研究室が公開している bert-for-japanese-whole-word-masking[v]を使用した。学習時のパラメータ設定は付録に記載する。

4.3 実験結果

前節で示した 4 通りの方法でファインチューニングした BERT について、4 通りのテストデータセットを用い、計 16 通りの評価を行った。テストデータセットは JSNLI および JSICK のテストデータセットに加え、前節で学習に使わなかった Wikipedia 入れ替えデータ約 14000 件と、新聞記事入れ替えデータ 10599 件の計 4 種類である。評価結果を表 5 に示す。

[iv] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[v] https://huggingface.co/cl-tohoku/bert-base-japanese-char-whole-word-masking

表 5　各テストデータに対する評価結果

A:accuracy, P:クラス毎の precision(矛盾/含意/中立), R:クラスごとの recall(矛盾/含意/中立)

学習データ		JSNLI	JSICK	Wikipedia shuffle	新聞記事 shuffle
JSNLI	A	91.93	50.9	0.17	0.12
	P	89.34/93.46/92.69	25.41/55.85/92.31		
	R	94.98/95.04/85.92	64.24/97.43/30.77		
JSICK	A	45.4	86.44	3.99	4.41
	P	51.85/91.94/38.34	78.55/75.92/93.85		
	R	1.21/32.68/97.59	84.57/90.99/85.31		
JSNLI+ Wikipedia shuffle	A	91.34	51.37	99.69	96.85
	P	89.05/93.26/91.33	26.94/56.01/93.26		
	R	93.59/94.69/85.77	69.64/96.42/30.47		
JSICK+ Wikipedia shuffle	A	39.22	87.94	98.96	92.34
	P	10.74/89.64/36.62	88.67/80.27/90.73		
	R	1.12/15.71/97.74	77.54/86.76/91.09		

含意関係認識データセットのみで学習した BERT では、名詞入れ替えデータでの正答率が著しく低い。提案手法である含意関係認識データセットと名詞入れ替えデータを合わせて学習した BERT は、JSNLI や JSICK テストデータの正答率はほぼ同等であるが、名詞入れ替えデータでの正答率が大幅に向上した。入れ替えデータは矛盾ラベルのみで構成されているが、バランスデータである JSNLI や JSICK でも高い性能を達成できていることから、単に矛盾のみを回答するのではなく純粋に性能が向上したといえる。

5 考察

Wikipedia でファインチューニングした場合に、名詞入れ替えによる矛盾ではなく、Wikipedia 特有の文体などを手掛かりに正答できた可能性があるが、新聞記事による評価も同等の高性能を示しているため、一般に入れ替えの結果となった矛盾関係を認識できるようになったと考えられる。事前学習データが Wikipedia であるため、既知の文であるために性能が上がった可能性についても、同様に新聞記事が高性能であることから排除できる。

JSNLI や JSICK の学習データだけでファインチューニングした BERT が、Wikipedia の入れ替えデータと新聞記事の入れ替えデータにおいて、推論したクラスの数を表 6 に示す。JSNLI や JSICK だけで学習した BERT は入れ替えデータでほとんど含意と推論していたことが分かる。これは、従来手法の学習では文脈よりも単語に重みづけした推論を行っているからではないかと考えられる。

入れ替えデータも含めて学習した BERT が、Wikipedia や新聞記事の名詞入れ替えデータにおいて不正解となっていた例を表 7 に示す。例のように格助詞「の」の前後で名詞が入れ替わる例が多く、その場合「日本の東京」⇔「東京の日本」など文意が変わるものが多い一方で、表 7 の例や表 3 で示した「文意が変わらない」例のように文意が変わらないものが入れ替えデータの自動生成の結果に含まれており、そのため判別できなかった可能性がある。

表 6　推論したクラスの数

学習データ	Wikipedia shuffle			新聞記事 shuffle		
	矛盾	含意	中立	矛盾	含意	中立
JSNLI	24	13804	56	13	10558	28
JSICK	555	13321	8	467	10126	6

表 7　名詞入れ替えデータでの不正解の例

前提	全国の高等専門学校の学生が参加する行事。
仮説	高等専門学校の全国の学生が参加する行事。

6 おわりに

本研究では、名詞を入れ替えた日本語文章データを作成し、既存の含意関係認識データセットと合わせて BERT をファインチューニングした。その結果、従来の含意関係認識の性能を保持しつつ、文内で名詞を入れ替えたときにも、意味が変わり矛盾になったかどうか判別できるようになった。今後は、文章内で名詞が入れ替わっても意味が変わらない文に対応することや、より長い文でも性能が保てているかをテストし、本研究の提案手法を応用してきたい。

謝辞

本研究は JSPS 科研費 JP22H00804, JP21K18115, JP20K20509, JST AIP 加速課題 JPMJCR22U4, およびセコム科学技術財団特定領域研究助成の支援をうけた。中日新聞社より、過去の記事データをご提供いただいた。ここに深謝申し上げる。

参考文献

[1]. 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 情報処理学会 第 244 回自然言語処理研究会, 2020.

[2]. 谷中瞳, 峯島宏次. JSICK: 日本語構成的推論・類似度データセットの構築. 人工知能学会 第 35 回人工知能学会全国大会, 2021.

[3]. 林部祐太. 知識の整理のための根拠付き自然文間含意関係コーパスの構築. 言語処理学会 第 26 回年次大会, 2020.

[4]. Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, Koichi Takeda. Overview of NTCIR-9 RITE: Recognizing Inference in TExt. Proceedings of NTCIR-9 Workshop Meeting, 2011. 291-301.

[5]. Hideo Joho, Tetsuya Sakai. Overview of NTCIR-10. Proceedings of the 10th NTCIR Conference, 2013.

[6]. Hideo Joho, Kazuaki Kishida. Overview of NTCIR-11. Proceedings of the 11th NTCIR Conference, 2014.

[7]. 小谷通隆, 柴田知秀, 中田貴之, 黒橋禎夫. 日本語 TextualEntailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識. 言語処理学会　第 14 回年次大会, 2008. 1140-1143.

[8]. Samuel R. Bowman, Gabor Angeli, Christopher Potts, Christopher D. Manning. A large annotated corpus for learning natural language inference. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015. 632-642.

[9]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 2019. 4171-4186.

[10] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014. 216-223.

[11]. Ai Ishii, Hiroshi Miyashita, Mio Kobayashi, Chikara Hoshino. NUL System at NTCIR RITE-VAL tasks. Proceedings of the 11th NTCIR Conference, 2014. 249-254.

[12]. 河原大輔, 黒橋禎夫. 高性能計算環境を用いた Web からの大規模格フレーム構築. 情報処理学会研究報告, 2006. 67-73.

[13]. 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理 14 巻 4 号, 2007. 67-81

A 付録

BERT の学習時設定において、最適化手法は AdamW[vi]を使用した。学習時のハイパーパラメータは以下の通りである。

・Epoch 数：10
・学習率：2e-5
・バッチサイズ：16

[vi] https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html