手がかり表現に基づく非論理的な言語推論の学習

張 辰聖子 [1]　持橋 大地 [2]　小林 一郎 [1]
[1] お茶の水女子大学　[2] 統計数理研究所
{g1920524,koba}@is.ocha.ac.jp　daichi@ism.ac.jp

概要

人間が行っている推論は論理的な含意関係だけではなく, 日常的な知識を背景にした常識推論が大きな役割を果たしている. 本研究はこの際の知識を, 複雑な内容を表現できる自然言語文自体ととらえ, 「から」「ため」のような手がかり表現を利用することで, 前提と帰結がともに文となる自然言語推論をコーパスから深層学習モデルとして直接学習する. コーパスから抽出した 157 万文ペアについて学習し, テストデータにおいて, 前提から生成した帰結文について人手評価を行ったところ, 65.8%が妥当な推論であるとの結果を得た. さらにエラーの原因について考察し, 改善と今後の可能性について議論する.

1　はじめに

ロボットのような人工知能が人間と共生する未来社会において, 人間と同等の推論や知的判断が行える方法を開発することは, 重要な研究課題だと考えられる. 自然言語処理の最近の推論の研究は含意関係認識のような論理的な推論を扱っていることが多いが, これでは扱えない自然な「推論」も実際には重要である. たとえば, 富士山と湖の写った写真を見て「富士山がある」「湖が見える」と判断するだけでなく, そこから「キャンプによい」「日本的な風景だ」と判断することは論理的な推論ではないが, 人間が常に行っている常識的な推論である. そこで本研究では, 「〜ため」のような手がかり表現とニューラル言語モデルを用いることで, 大量のコーパスから直接, こうした常識的推論を学習する. 提案法の特徴は, 前提と帰結がどちらも自然言語の文となる言語的推論であり, 従来のように論理構造を介していないことにある.

2　関連研究

こうした非論理的な[1)]推論は, 自然言語処理では常識推論 (commonsense reasoning) とよばれて長く議論されており, 特に最近活発に研究されている分野である [1]. ただし, このためのデータセット [2][3] がそうであるように, 従来はこうした知識はあくまで知識グラフのような離散的な知識構造で表されていた. こうしたグラフの作成やラベル付けには大量の人手が必要である上, 多くは単語間のグラフ関係であるため, 複雑な知識を表すことができず, 表現力に乏しいという大きな欠点がある.

いっぽう, 推論についてこうした知識構造を介さず, 文から直接文を導く自然言語推論が行われ始めている [4]. こうした自然言語推論は論理的含意を対象にしているが, この方法は常識推論にも自然に適用でき, 本研究ではその可能性を模索する. Yu ら [5] は, 短い文のペアを人手でラベル付けしたデータセット ASER [6] から, 文の連続をグラフ上のランダムウォークとしてサンプリングし, そのデータに基づく常識推論の学習を提案している. ASER は大量のコーパスからあらかじめ作成されたものであるが, 本研究は関係を絞って, 常識推論を直接コーパスから学習するものであるといえる.

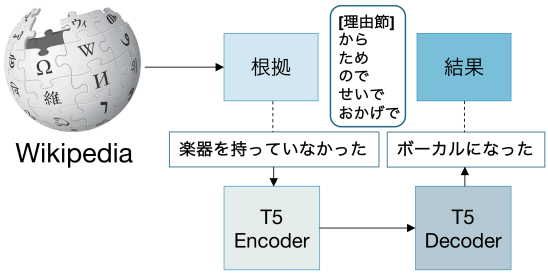

図 1　Wikipedia からのデータ生成と，推論の学習の概要.

3　自然言語による推論

3.1　研究概要

図 1 に研究の概要を示す．日本語 Wikipedia コーパスから，「理由」を表現する手がかり表現を元に

1)　論理的推論と区別するために, あえて「非論理的」という言葉を使っているが, これは推論に結果として論理的推論が含まれることを排除するものではない.

根拠と結論がペアになった文を抽出する．根拠部分を入力，結論部分を出力とし，文を生成する深層学習モデルである T5 に学習させる．これにより根拠に相当する文から結論を示す文を生成することを通じて, 自然言語推論を実現する．

3.2 データ作成

データの収集には日本語版 Wikipedia コーパスを用いた．以下にデータ作成の手続きを示す．

Step 1. 正規表現で理由節を持つものを抽出 日本語版 Wikipedia 本文から，理由節を持つ 1 文を正規表現を用いて抽出する．その際，手がかり表現として「から」「ので」「ため」「おかげで」「せいで」の 5 つを用いた．また，理由節を持つ文の文頭が指示詞（「この」,「その」）であった場合，前文で置換をするという簡単な指示詞の補完を行った．

Step 2. 形態素解析によるフィルタリング 正規表現によって抽出された理由節を持つ文に対し，1 文ごとに MeCab を用いて形態素解析を行う．正規表現での抽出では，5 つの手がかり表現は，理由を表す以外の意味を持つことがある．そこで，形態素解析により主に理由を表す品詞として使われている場合のみを抽出する．以下に品詞による使われ方の違いの例を示す．

- 信号が青から赤に変わった．(格助詞, 変化の「から」)
- 楽器を持っていなかったからボーカルになった．(接続助詞, 理由の「から」)

例に挙げたように「から」の品詞細分類が接続助詞である時，主に理由を表すことから解析により抽出を行う．上記により，解析の結果，理由節が以下の形態素として含まれていた場合に，抽出を行う．

- から　助詞, 接続助詞
- ので　助詞, 接続助詞
- ため　名詞, 非自立, 副詞可能
- せい　名詞, 非自立, 一般
  ＋で　助詞, 格助詞, 一般
- おかげ　名詞, 一般
  ＋で　助詞, 格助詞, 一般

Step 3. 機械学習によるフィルタリング 形態素解析によって残った文のうち，名詞で用いられる「ため」はいくつかの意味を持つ．そのため，理由を表す「ため」のみを抽出することを目的として，訓練済み日本語 BERT モデルを用いて二値分類を行った．ここで用いた BERT モデルは Hugging Face[1] の自然言語ライブラリ Transformers[2] に基づく，東北大学公開の事前学習済み日本語 BERT モデル[3] を使用した．「ため」を手がかりとして抽出した文から，ランダムに 300 文を抜き出し，手動で理由を表すかそうでないかのタグ付けをし，二値分類の機械学習を行くった．300 文での 5 分割交差検証では 88.7 % の精度で理由を表すかそうでないかの分類を行えていることがわかった．分類を行った結果, 1,288,304 文の「ため」を含む文から，730,133 文 (56.7%) の理由を表す文を抽出することができた．

表 1 訓練データの例と，それに類似する前提を与えたときのファインチューニングを行った T5 の文生成出力例.

訓練データ 1	前提	小波まで漢字だと固いイメージになる
	結果	こなみとひらがなにした
訓練データ 2	前提	漢字ではイメージが固い
	結果	みずいろと平仮名にした
新しい入力	前提	漢字ではとっつきにくい
	出力	ひらがなにした

Step 4. 根拠表現と結果表現の抽出 日本語自然言語処理ライブラリである GiNZA[4] を用いて，抽出した文に対して係り受け解析を行う．坂地ら [7] の手法を用いて, 根拠表現と結果表現の抽出を行った．

3.3 推論モデル

収集した文ペアを日本語 T5 モデルに与え, 前提から結果を生成する深層学習モデルを学習することで，自然言語による推論を実現する．

4 実験

4.1 実験設定

本研究では，3.2 節で収集した 1,572,956 件の根拠と結果のペア文をデータとして用いる．実験の際は，このデータを訓練：開発：評価=0.95:0.025:0.025 として評価を行う．評価は評価データからランダムで 100 文を抽出し手動で行い，それに伴い

表 2 T5 モデルのファインチューニングの実験設定.

max input_len	max output_len	train batch_size	num train_epoch	learning rate
128	128	8	10	3e-4

1） https://huggingface.co
2） https://github.com/huggingface/transformers
3） https://huggingface.co/cl-tohoku/bert-base-japanese
4） https://megagonlabs.github.io/ginza/

表 3　評価データに対するファインチューニングを行った T5 の文生成出力例と，出力に対する人手評価の例.

分類の種類	文章
例 1　妥当な推論	前提　一部踏み跡程度の不明瞭な登山道や作業道などがある
	出力　通行には注意を要する 正解　登山の際には注意が必要である
例 2　妥当な推論	前提　最寄の鉄道駅から距離が遠い
	出力　自転車通学が多い 正解　本校は鉄道を利用して通学する生徒は少ないと思われる
例 3　前提の繰り返し	前提　本系列は荷重や引張力をタンク体で直接受ける構造である
	出力　荷重や引張力をタンク体で直接受ける構造であるため 正解　タンク体の強度が低下すると車両としての使用ができなくなる
例 4　矛盾	前提　温泉権は温泉が存在する土地の所有者に対して永年利用を認めさせる債権である
	出力　温泉権者に対して永年利用を認めさせることはできない 正解　土地の所有者がこれを取り除くためには少なくとも永年の利用に相当する補償が必要となる

BERT-Score[8] によって，生成文と正解文との意味的類似度を検証した．評価指標は 4.2 節に示す．本研究で用いる T5 モデルは Hugging Face[5] の自然言語ライブラリ Transformers[6] に基づく，Isao Sonobe 公開の事前学習済み日本語 T5 モデル[7] を使用した．

4.2　評価基準

表 3 の生成例 2 にあるように，評価データの正解と比べると全く違う出力であるが，出力内容自体は前提から妥当に導かれるものが多くあった．したがって，生成データセットにおけるモデルの性能は，出力の妥当性を人手で判断することで評価する．評価基準は，自然言語による演繹的な文生成を行うモデルの提案をしている関連研究 [4] を参考にし，以下の 6 つの分類とする．

i 妥当な推論である　出力は前提から妥当に導かれる結論である．

ii 妥当な推論であるが文法的な間違いがある　出力は前提から導かれうる結論だが，主語などの成分が欠けていたり，動詞の活用に違和感があるなど内容の理解を妨げない程度の文法的な間違いがある．

iii 前提を繰り返す　出力は，前提文の繰り返し，もしくは言い換えであり前提と同意である．または，前提に含まれる単語を組み合わせているだけである．

iv 前提から導かれない　出力は文，内容は正しいが，前提から導かれ得る結論ではない．

v 矛盾している　出力は，前提と矛盾する，もしくは本質的に間違っている．

vi 結論が理解し得ない　結論が文をなしていない．もしくは前提が文をなしていない．

表 4　人手評価されたデータに対する生成文と正解文の BERT-Score の平均.

種類	件数	Precision	Recall	F_1
full	100	0.678	0.700	0.688
i	52	0.699	0.719	0.707
ii	20	0.638	0.667	0.651
iii	11	0.688	0.695	0.690
iv	9	0.665	0.668	0.666
v	5	0.700	0.721	0.710
vi	3	0.597	0.654	0.623

4.3　実験結果

実験でファインチューニングされたモデルに評価データを与えると，表 3 のように文生成が行われた．その他の推論結果は, 付録 A の表 7 に示した. 前提はモデルに与えた文，出力はモデルから生成された文，正解は収集データ中の前提に対する結論文のことである. 生成データセットにおけるモデルの性能に対する人手で行った評価結果は，各評価者ごとの評価数を表 5 に，全ての評価者の平均をグラフに

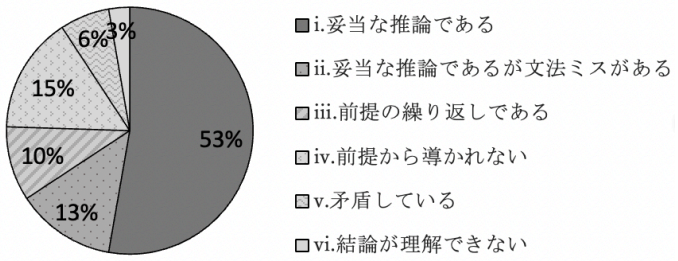

図 2　評価データからランダムに抽出した 100 文に対する人手評価の結果 (平均)

5)　https://huggingface.co
6)　https://github.com/huggingface/transformers
7)　https://huggingface.co/sonoisa/t5-base-japanese

表 5 評価データからランダムに抽出した 100 文に対する人手評価の結果 (件数).

種類	評価者 1	評価者 2	評価者 3	評価者 4	割合 (%)
i	52	47	45	67	52.8
ii	20	19	13	0	13.0
iii	11	10	9	9	9.8
iv	9	18	21	13	15.3
v	5	5	9	7	6.5
vi	3	1	2	4	2.8

したものを図 2 に示す．人手で評価したデータに対する，生成文と正解文の BERT-Score [9] による類似度を表 4 に示す．また，訓練データにあるような形の前提であれば妥当な推論を行えるという仮定の下，訓練データの例と，それに伴うモデルによる出力例を表 1 に示す．出力が妥当な推論である，もしくは妥当な推論であるが文法的な間違いがあると評価された割合は 65.8 %となった．

4.4 考察

前提を与えることで，文生成の形で推論を行うことができた．BERT-Score による文の類似度（表 4）を見ると，妥当な推論が行われている場合（評価 i）は他の評価指標の結果に比べ，類似度の平均が高くなっていることがわかる．しかし，評価 v の類似度の平均が最も高くなっている．これは，生成文が表 3 の例 4 のように，矛盾はしているが矛盾理由が文末尾による否定によるものである場合に，文全体の類似度自体は上がってしまったからだと考えられる．

エラーについて，3 つ目の評価指標である前提を繰り返すことに注目すると，表 3 の例 2 のように，前提を繰り返したのちに「ため」をつけるという事例がいくつかあった．理由としては，訓練データに結論が「ため」で終わる文が，4,293 件と多く含まれていたことが考えられる．そのようなデータが含まれてしまう原因は 2 つほど考えられる．

1 つ目は，「ため」によって収集された文に「のために していたため」といった表 6 の例 1,2 に示したように「ため」を多重で使った文が多くあったことが原因だと思われる．このような「ため」多重文に対し，データ生成における Step 4 を行うと，結果部分が「ため」で終わるデータが抽出されてしまう．また，3.2 節のデータ作成で説明したように「ため」はいくつかの意味を持つ．その中でも，「ため」多重文で用いられる「ため」は例 2 のように「目的」の意味で使われていることが多くあり，「ため」に対する機械学習の精度の向上が望まれる．

表 6 エラー分析の結果，原因であると考えられる訓練データのノイズ例. 元の文と, 抽出した文ペアを示した.

例 1 子供を増やすことは世帯所得増加に繋がるため， 子供を産むインセンティブとなるために少子化 対策となりうるという考えがある．
前提　子供を増やすことは世帯所得増加に繋がる 結果　子供を産むインセンティブとなるため
例 2 彼らに対抗するために資金を調達するために ツアーに行くと，彼らはギャングによって 絶えず脅かされています．
前提　彼らに対抗する 結果　資金を調達するため
例 3 必要以上の事件を寄せ付けないためにクラス A の 身分を隠し，小さな探偵事務所を構える悪行の 元に，今日もまた事件がやってくる．
前提　必要以上の事件を寄せ付けない 結果　必要以上の事件を寄せ付けないため

2 つ目の原因としては，訓練データの結果の中に，表 6 の例 2 のように前提を繰り返した後に「ため」をつけたものもあった．こうしたことが起こった原因は，係り受け解析時に「ため」の前で文が区切られてしまっていたからである．例 2 の文では，「必要以上の事件を寄せ付けないために」で区切られてしまい，データ生成における Step 4 を行うと, 前提を繰り返す文を結果として抽出してしまう．これは，係り受け解析器を用いて自動的に前提と結果を抽出しているため起こりうるエラーである．

5 おわりに

本研究では，自然言語文そのままの形で推論を行う手法の開発を目的に，自然言語による推論を自然言語文生成として表現する研究に取り組んだ．日本語 Wikipedia コーパスから，理由節を手がかり表現として根拠と結論を抽出した. 抽出した因果関係のデータを用いて，根拠部分を入力，結論部分を出力として深層学習を行うことで，推論を行う形の文生成が可能なことがわかった．モデルによって生成された文に対して人手で評価を行った結果，前提から導かれる推論として妥当だと言える割合は 65.8%となった．今後は，発生したエラーを改善して収集データのノイズを減らし，推論文生成の精度を上げたい．また，作った推論モデルを用いて，多段推論および前提を予測する帰納的推論などに挑戦していきたいと考えている．

参考文献

[1] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense Reasoning for Natural Language Processing. In *ACL 2020*, pp. 27–33, 2020.

[2] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI 2017*, p. 4444–4451, 2017.

[3] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, and Hannah Rashkin. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI 2019*, p. 3027–3035, 2019.

[4] Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, and Greg Durrett. Flexible Generation of Natural Language Deductions. In *EMNLP 2021*, pp. 6266–6278, 2021.

[5] Changlong Yu, Hongming Zhang, Yangqiu Song, and Wilfred Ng. CoCoLM: Complex commonsense enhanced language model with discourse relations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1175–1187, 2022.

[6] Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. ASER: A Large-scale Eventuality Knowledge Graph. In *WWW 2020*, p. 201–211, 2020.

[7] 坂地泰紀. 構文パターンを用いた因果関係の抽出. 言語処理学会第 14 回年次大会論文集, 2008, pp. 1144–1147, 2008.

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, Vol. abs/1904.09675, , 2019.

[9] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020.

付録 A 推論結果の例

表 7　学習した T5 深層学習モデルによる推論結果の例.

	分類種類		文章
例 1	妥当な推論である	前提	戦争支持の風潮に背を向けた
		出力	批判された
		正解	非国民のレッテルを張られる
例 2	妥当な推論である	前提	レーダー類が貧弱である
		出力	対艦攻撃能力は低い
		正解	射撃管制手段は不明だ
例 3	妥当な推論である	前提	中々仕事がない
		出力	暇をもてあます
		正解	バイトをしている
例 4	妥当な推論である	前提	帰任先が大阪だった
		出力	関西弁が話せる
		正解	現在大阪在住
例 5	妥当な推論であるが文法的な間違いがある	前提	騎兵隊という部隊名が付いているが馬の輸送が間に合わずまた連発銃の普及により騎馬のメリットが薄れていた
		出力	格下げされた
		正解	実際は兵士は徒歩だったルーズベルトは黒馬に乗っていた
例 6	妥当な推論であるが文法的な間違いがある	前提	光化学系や光化学系の還元力を蓄積させ活性酸素を発生させる
		出力	触媒とし
		正解	植物にとって害となる
例 7	前提を繰り返す	前提	e は超越数であると証明されている
		出力	e は超越数であると証明されているのだ
		正解	背理法によりは無理数でなければならない
例 8	前提を繰り返す	前提	マタンサ自体が普通名詞である
		出力	スペイン語圏ではマタンサは普通名詞である
		正解	誤解を防ぐため文章化に当たっては atanzadelcerdo ブタのマタンサなどと表記する
例 9	前提から導かれない	前提	幼い頃から弟の面倒を見てきた
		出力	家計は裕福だ
		正解	子供の面倒を見る
例 10	前提から導かれない	前提	咽喉頭の異常感は古代ギリシア時代から記載があるといわれておりヒポクラテスは更年期の女性が起こす病態と捉え当時子宮は体内を移動する臓器と考えられていた
		出力	ヒポクラテスは子宮の異常感を女性特有の生理現象と捉えた
		正解	これが喉につかえる感じを生じさせていると考えた
例 11	矛盾している	前提	生息数が減少している可能性が高い
		出力	絶滅のおそれは低いと考えられている
		正解	C は絶滅危惧種としている
例 12	矛盾している	前提	組織内に寄生したプレロセルコイドに有効な駆虫薬はない
		出力	駆虫薬を併用する
		正解	外科的に摘出することが有効である
例 13	結論が理解し得ない	前提	1 編成を 1 車両として運用する考え方を基本とした
		出力	これ
		正解	連結器は中間連結部は半永久連結器を使用している
例 14	結論が理解し得ない	前提	1 の値と独立ではない
		出力	1
		正解	末尾事象ではない