ニューラル自然言語推論に向けた Monotonicity に基づく含意関係認識データセット自動構築

谷中 瞳[1,2,a] 峯島 宏次[2] 戸次 大介[2] 関根 聡[1]
乾 健太郎[1,3] Lasha Abzianidze[4] Johan Bos[4]
[1] 理化学研究所 AIP センター [2] お茶の水女子大学 [3] 東北大学 [4]University of Groningen
{hitomi.yanaka[a], satoshi.sekine}@riken.jp, mineshima.koji@ocha.ac.jp,
bekki@is.ocha.ac.jp, inui@ecei.tohoku.ac.jp, {l.abzianidze, johan.bos}@rug.nl

1 はじめに

自然言語推論 (Natural Language Inference, NLI)，または，含意関係認識（Recognizing Textual Entailment, RTE）[1] とは，自然言語理解を目的としたタスクの 1 つである．このタスクでは，前提文 T が仮説文 H を含意するか否かを自動判定する．近年，含意関係認識においてニューラルネットワークを用いた手法 [2, 3] が活発に研究されており，モデルの学習・評価には，SNLI [4] や MultiNLI [5] といった，クラウドソーシングによって構築された大規模な含意関係認識データセットが用いられている．しかし，これらのデータセットは前提文と仮説文の両方を学習データに含めなくても，仮説文のみを学習させることによってベースラインをはるかに上回る正答率が得られてしまう，という問題が指摘されている [6, 7, 8]．また，SNLI や MultiNLI は量化や否定，比較表現，条件文といった，実世界のテキストに現れる多様な推論現象を網羅していないことも指摘されており [9]，ニューラル含意関係認識モデルが論理語や機能語の意味を学習できているかは自明ではない．

一方で，推論現象に基づく含意関係認識データセットとして，FraCaS [10] がある．このデータセットは言語学者のチームによって設計・構築されており，幅広い推論現象を扱っている．しかし，語彙知識は推論に最低限必要な知識に制限されている．また，データ数は 346 件と小規模であり，ニューラルモデルの学習データとして用いることは想定されていない．

多様な語彙知識や統語構造からなる実世界のテキストから論理推論のデータセットを構築するためには，高度な言語学・論理学の知識を必要とする．そのため，クラウドソーシングによる構築は困難であり，また，言語学者が大量の実テキストを一件ずつ分析し構築することも現実的ではない．そこで本研究では，論理推論の一つである monotonicity に着目して，多様な語彙知識と統語構造における含意関係認識データセットを自動構築する手法を提案する．構築したデータセットを最高精度のニューラル含意関係認識モデルの学習データに追加すると，一部の推論現象を除き，monotonicity の推論に関する含意関係認識の問題において実際に正答率が向上した．また，正答率の向上がそれほど顕著ではない推論現象についても，その理由について分析を行った．なお，構築した含意関係認識データセットは研究利用が可能な形式で公開する予定である．

2 先行研究

論理推論のデータセットを自動構築する先行研究として，Bowman ら [11] は簡潔な人工言語の文法規則に基づいて量化子を含む前提文と仮説文のペアを自動生成する手法を提案した．Geiger ら [9] はこの手法を発展させて，自然言語の文法規則に基づいて多重量化 (multiple quantification) を含む前提文と仮説文のペアを自動生成する手法を提案した．以下に Geiger らの手法によって自動生成された，多重量化を含み含意関係が成立する文ペアの例を示す．

T: Not every ugly dwarf does not licks some rifle
H: Every ugly dwarf does not tediously licks no Canadian rifle

これらの手法では，1 つの構文フレームに対してランダムに有限個の語彙を当てはめ，文ペアを自動生成する．そのため，上の例のように生成文の統語構造は単文に固定化され，実テキストにあるような，多様な統語構造からなる文は生成できない．また，Geiger らの手法では上位－下位などの語彙関係は扱っておらず，Bowman らの手法では語彙数が数個に限られている．

語彙関係については，既存の含意関係認識データセットに対して，いくつかの構文ルール [12, 13] や述語項構造 [14] に基づいて語彙の入れ替え箇所を特定し，外部辞書等を用いて語彙を入れ替えることで，語彙関係を問う問題を増やす手法が提案されている．しかし，これらの手法では，論理語や機能語との関係については考慮していない．したがって，多様な語彙知識と統語構造を扱う論理推論のデータセットを自動構築する手法は存在しない．しかし，そのようなデータセットがあれば，ニューラルモデルが自然言語の推論を学習するためには何が必要なのかという根源的な問いに対して，一つの解を提示することが期待できる．そのようなデータセットを自動構築するためには，実

テキストから各語の語義と統語構造，そして推論の性質を考慮してデータを構築する必要がある．

3 背景

monotonicity は語の入れ替えに基づく推論であり，量化子などの表現が持つ monotonicity の性質に基づいて文中の語を入れ替えた文と，元の文との含意関係が成り立つという推論である [15, 16]．polarity は monotonicity の性質に応じて決定する文中の語の性質であり，+ が付与されている場合，その語を意味的に上位の表現に置き換えても，元の文と含意関係が成り立つ．また，− が付与されている場合は，その語を意味的に下位の表現に置き換えても，元の文と含意関係が成り立つ．ある表現が upward monotone である (項関係にある語の横に↑を付与して表す) とは，その表現と項関係にある語の polarity を維持するという意味である．逆に，ある表現が downward monotone である (項関係にある語の横に↓を付与して表す) とは，その表現と項関係にある語の polarity を反転させるという意味である．また，ある表現が non-monotone であるとは，upward monotone, downward monotone のいずれでもないことを指す．

一般化量化子は monotonicity の推論と関連が深く，一般化量化子の monotonicity の性質に応じて，その一般化量化子をもつ名詞句 (NP), その名詞句が項となる動詞句 (VP) の polarity が決定する．例として，*Some boys are gracefully dancing* という文を考える．一般化量化子の一つである *some* は名詞句，動詞句に対し upward monotone である．このため，(1) に示すように，*boy* と *are gracefully dancing* の polarity はそれぞれ + となり，これらを上位語に置き換える，修飾語を除く，or で結ぶ等位接続表現を導入する，といった操作によって，上位の表現に置き換えても，元の文と含意関係が成り立つ．

(1) Some [NP boys^{+}]↑[VP are gracefully dancing^{+}]↑
⇒ Some [NP people] [VP are dancing]
⇒ Some [NP boys] [VP are gracefully dancing or singing]

逆に，*boy* と *are gracefully dancing* を下位語に置き換える，修飾語を加える，and で結ぶ等位接続表現を導入する，といった操作によって，下位の表現に置き換えると，(2) に示すように，元の文との含意関係が成り立たなくなる．

(2) Some [NP boys^{+}]↑[VP are gracefully dancing^{+}]↑
⇏ Some [NP tall boys] [VP are very gracefully dancing]
⇏ Some [NP schoolboys] [VP are dancing and singing]

no などの否定表現は名詞句，動詞句に対し downward monotone である．そのため，(3) に示すように *boy*，*are gracefully dancing* の polarity は − となり，下位の表現に置き換えた文は元の文と含意関係が成り立ち，上位の表現に置き換えた文は元の文と含意関係が成り立たなくなる．

(3) No [NP boys^{-}]↓[VP are gracefully dancing^{-}]↓
⇒ No [NP tall boys] [VP are gracefully dancing]
⇏ No [NP people] [VP are dancing]

否定・条件節もまた downward monotone であり，否定・条件節のスコープ内にいる語の polarity は (4) に示すようにさらに反転する．

(4) If [there are no [NP boys^{+}]↓[VP gracefully dancing^{+}]↓]↓, [the event might be canceled]↑
⇒ If [there are no [NP people] [VP dancing]], [the event might be canceled]

このように，語の polarity は周辺環境の monotonicity の性質に応じて反転することがあるため，統語構造から慎重に語の polarity を計算する必要がある．

4 データセットの構築

4.1 使用コーパス

本研究では，統語・意味解析情報つき多言語コーパスである Parallel Meaning Bank (PMB) [17] を用いて含意関係認識データセットを自動構築する．PMB を用いる理由は 3 点ある．一つには，PMB には統語・意味解析情報が付与されており，これらの情報を用いて効率的にデータセットを自動構築できることである．具体的には，PMB に付与されている語義タグは，monotonicity に基づく上位・下位の表現への自然な置き換えを容易にする．また，PMB には Combinatory Categorial Grammar (CCG) [18] に基づく統語解析結果と意味現象タグ [19] が付与されており，monotonicity と polarity の特定に役立つ．二つ目の理由は，PMB は様々なジャンルの文から構成されており，語彙的・構文的に多様な文を含んでいることである．三つ目の理由は，PMB の一部（Gold・Silver）は人手で統語・意味解析情報が付与されているため，自動構築されたデータセットの品質を保証してくれることである．提案手法では簡単な推論のデータが構築されてしまうことを防ぐため，Gold・Silver の英文のうち，5 トークン以上からなる平叙文約 6 万 4 千件を対象とした．

4.2 提案手法

3 節で述べたように，monotonicity の推論を扱うデータを構築するためには，語彙が持つ monotonicity の性質と統語構造に基づいて文中の語の polarity を決定し，適切に語彙を上位・下位の表現に置き換える必要がある．そこで提案手法ではまず，PMB にある文から monotonicity の推論に関わる量化表現または等位接続表現 (and, or) を含む文を意味現象タグ[1]を用いて特定する．例として *We have consumed all the natural resources* という文には下記のように意味現象タグが付与されている．

We	have	consumed	**all**	the	natural resources
PRO	NOW	EXT	AND	DEF	CON

[1] AND (*all*, *every*, *each*, *and*), DIS (*some*, *several*, *or*), NEG (*no*, *neither*), DEF (*both*), and QUV (*many*, *few*) を対象とした．

表 1: 構築したデータセットの例．∗が付いている文は PMB にある文である．

現象	件数	前提文	仮説文	正解
Up	7784	*Tom asked Mary to go to the store to buy some caraway seed bread*	*Tom asked Mary to go to the store to buy some bread**	entailment
Down	21192	*We have to abolish all nuclear weapons, because they are deadly to mankind**	*We have to abolish all atomic bombs, because they are deadly to mankind*	entailment
Conj	6076	*If I had time, I'd travel to Europe*	*If I had time and money, I'd travel to Europe**	entailment
Disj	438	*The trees are barren*	*The trees are barren or bear only small fruit**	entailment

表 2: 評価結果．（）内の数字は各評価データの件数を示す．

学習データ	構築した評価用データ (corr %)					GLUE (corr %)						FraCaS	SICK
	Up (100)	Down (100)	Conj (100)	Disj (100)	All (400)	Up (30)	Down (26)	Non (22)	Conj (24)	Disj (22)	Total (124)	corr % (80)	acc % (4927)
MNLI	34.8	-42.7	75.8	-13.9	8.9	50.4	-67.5	23.1	52.5	-6.1	17.8	42.2	55.4
MNLI+[Geiger+ 2018]	26.6	1.3	73.8	-14.0	19.4	59.6	-49.3	14.0	62.1	-18.8	26.3	45.7	58.2
MNLI+構築したデータ	**72.7**	**88.0**	**81.9**	**-7.8**	**43.0**	**67.0**	**29.8**	**47.9**	**72.1**	**-4.1**	**51.2**	**48.7**	**60.0**

次に，CCG に基づく統語解析結果から量化子をもつ名詞句と動詞句を特定し，それぞれの polarity を決定する．なお，意味現象タグ `NEG`,`IMP` を用いて否定や条件節の有無を特定し，対象となる名詞句と動詞句が否定や条件節のスコープ内に存在する場合には，3 節で述べたように polarity を反転させる．

We have consumed **all** [$_{\mathsf{NP}}$ **the natural resources**$^{-}$]↓

次に，polarity に応じて，名詞句・動詞句をトークンに付与されている語義タグと WordNet [20] を用いて上位・下位の表現に置き換え，元の文と含意関係が成り立つ仮説文を生成する．トークンに語義タグが付与されていない場合は，Lesk [21] を用いて語義を特定し，上位・下位の表現に置き換える．また，上位の表現への置き換えは修飾語の削除によっても行う．この操作によって，以下のように含意関係にある前提文と仮説文のペアが生成できる．

T: We have consumed all [$_{\mathsf{NP}}$ the **natural** resources^{-}]↓
H: We have consumed all [$_{\mathsf{NP}}$ the **mineral** resources]

最後に，以下のように前提文と仮説文を交換することで，含意関係が成立しない文ペアが生成できる．

T: We have consumed all the mineral resources
H: We have consumed all the natural resources

4.3 構築したデータセット

提案手法によって，前提文と仮説文のペア合計 3 万 6 千件を構築した．表 1 に構築したデータの例を示す．正解ラベルの分布は entailment/neutral = 49%/51% であった．語彙数は 1 万 5 千件であった．構築したデータのうち，現象ごとにランダムに選んだ 100 件ずつを評価用データとし，残りを学習用データとした．

5 実験と評価

5.1 実験設定

最高精度のニューラル含意関係認識モデルの一つである BERT [2] を用いて，学習データによって含意関係認識の精度が変わるのかについて，評価を行った．

学習データ MultiNLI (39 万件)，MultiNLI+多重量化を含むデータセット [9] (89 万件)，MultiNLI+本研究で構築したデータセット (42 万件) の 3 種類の学習データを用意した．

評価データ 構築した評価用データ，GLUE [3] (upward monotone, downward monotone, non-monotone, conjunction, disjunction)，FraCaS(一般化量化子のセクション），SICK 評価用データ [22]，という monotonicity の推論を扱う 4 種類の評価データを用いて，精度を評価した．SICK による評価では，正答率を評価指標に用いた．それ以外のデータは件数が少なくラベルの分布に偏りがあるため，GLUE の評価方法に従い，Matthiews 相関係数（$[-1, 1]$ の範囲の値を取りうる）を評価指標に用いた．

5.2 実験結果

表 2 に評価結果を示す．構築したデータセットを学習データに追加することで，いずれの評価データにおいても精度が向上した．多重量化を含むデータセットとの比較では，構築したデータセットを追加した場合の方が，学習データのサイズは小さいのにもかかわらず，より精度が向上した．2 節で例を示したように，多重量化を含む論理推論のデータセットは monotonicity の推論に特化したデータセットではないため，monotonicity の推論を扱う問題における正答率の向上に直接的にはつながらなかったと考えられる．このことは，学習データのサイズよりも，学習させたい能力に合わせて適切に学習データを設計することによって精度が改善する可能性を示唆している．SICK による評価結果について，MultiNLI の学習データでニューラル含意関係認識モデルを学習した場合は，SICK の学習データで訓練した場合よりも正答率が 40-50%と低くなることが報告されている [23]．本実験の結果から，BERT においても同様に正答率が 40-60%と低く，予測精度が学習データのドメインに依存することを示唆している．

5.3 分析

他の現象と比較して，disjunction では著しい精度向上は見られなかった．そこで，モデルがどのような問題で，学習データを変えても共通して誤答しているのか，SICK 以外の評価データ 604 件についてエラー分析を行った．次のような，仮説文の語は全て前提文中の語からなるが，含意関係が成り立たない例 64 件で誤答していた．

T: He is either in London or in Paris
H: He is in London

また，次のような，仮説文に前提文には現れない語が含まれているが，含意関係が成り立つ例 35 件で誤答していた．

T: I don't want to have to keep entertaining people
H: I don't want to have to keep entertaining people who don't value my time

このような推論は disjunction, downward monotone に特徴的である．構築したデータに disjunction, downward monotone の問題は約 2 万 1 千件含まれているのにもかかわらず誤答していることから，学習データの不足が誤答の原因ではない可能性があり，今後さらなる分析を行う．

6 おわりに

本稿では，monotonicity の推論の性質を用いて多様な語彙知識と統語構造における含意関係認識データセットを自動構築する手法を提案した．構築したデータセットを学習データに追加してモデルを学習することで，一部の推論現象を除き，monotonicity の推論に関する含意関係認識の問題において精度向上が見られた．今回は PMB に付与された情報を用いてデータセットを構築したが，CCG の統語解析器や意味現象タグ付与ツール [24] を用いることで，任意のデータから全自動で構築することも可能である．今後の展望として，全自動で構築したデータでも，データの品質を担保でき，同様の精度向上が見込めるのか検討する．

謝辞 本研究は JST，AIP-PRISM，JPMJCR18ZM の支援を受けたものである．

参考文献

[1] Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2013.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.

[3] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding, 2018.

[4] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP-2015*, pp. 632–642, 2015.

[5] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL HLT-2018*, pp. 1112–1122, 2018.

[6] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proc. of NAACL HLT-2018*, pp. 107–112, 2018.

[7] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proc. of the 7th Joint Conference on Lexical and Computational Semantics*, pp. 180–191, 2018.

[8] Masatoshi Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proc. of LREC-2018*, 2018.

[9] Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. Stress-testing neural models of natural language inference with multiply-quantified sentences, 2018.

[10] Robin Cooper, Richard Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. FraCaS–a framework for computational semantics. In *Deliverable*, Vol. D6. 1994.

[11] Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. Recursive neural networks can learn logical semantics. In *Proc. of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 12–21, 2015.

[12] Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking NLI systems with sentences that require simple lexical inferences. In *Proc. of ACL-2018*, pp. 650–655, 2018.

[13] Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proc. of the 2018 EMNLP Workshop BlackboxNLP*, pp. 337–340, 2018.

[14] Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. Ordinal common-sense inference. *TACL*, Vol. 5, pp. 379–395, 2017.

[15] Johan Van Benthem. Determiners and logic. *Linguistics and Philosophy*, Vol. 6, No. 4, pp. 447–478, 1983.

[16] Thomas Icard and Lawrence Moss. Recent progress in monotonicity. *LILT*, Vol. 9, No. 7, pp. 167–194, 2014.

[17] Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proc. of EACL-2017*, pp. 242–247, 2017.

[18] Mark Steedman. *The Syntactic Process*. MIT Press, 2000.

[19] Lasha Abzianidze and Johan Bos. Towards universal semantic tagging. In *Proc. of IWCS-2017*, pp. 1–6, 2017.

[20] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, 1998.

[21] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. of the 5th Annual International Conference on Systems Documentation*, pp. 24–26, 1986.

[22] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proc. of LREC-2014*, pp. 216–223, 2014.

[23] Aarne Talman and Stergios Chatzikyriakidis. Testing the generalization power of neural network models across NLI benchmarks, 2018.

[24] Johannes Bjerva, Barbara Plank, and Johan Bos. Semantic tagging with deep residual networks. In *Proc. of COLING-2016*, pp. 3531–3541, 2016.