ニューラルネットは自然言語推論の体系性を学習するか

谷中 瞳[1,2] 峯島 宏次[2] 戸次 大介[2] 乾 健太郎[1,3]
[1] 理化学研究所 [2] お茶の水女子大学 [3] 東北大学

hitomi.yanaka@riken.jp, mineshima.koji@ocha.ac.jp,
bekki@is.ocha.ac.jp, inui@ecei.tohoku.ac.jp

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，計算機による自然言語理解の実現に向けたタスクの 1 つであり，与えられたテキスト T から仮説 H が推論されるかどうかを自動判定するタスクである．近年，SNLI [2] や MultiNLI [3] といった，クラウドソーシングによる大規模な RTE データセットの発展に伴い，ニューラルネットワークを用いた手法 [4, 5] が活発に研究されている．しかし，SNLI や MultiNLI には仮説文だけで正解を予測できるバイアスが含まれるといった問題が指摘されており [6, 7, 8]，ニューラル RTE モデルがどの程度自然言語の推論に対して汎化性能を有するのかは自明でない．推論における汎化性能を示す概念として，推論の体系性 (systematicity) [9] がある．推論の体系性とは，人間の認知に関する 2 つの理論であるコネクショニズムと古典的計算主義の対立から生まれた概念であり，個々の規則が既知であれば，それらの組み合わせについても推論可能であるという認知の性質を示す．

本研究では，重要な推論現象の一つである monotonicity [10] に焦点を当てて，ニューラル RTE モデルが推論の体系性を捉えられているかについて分析する．monotonicity とは，文の構成素の極性（上方含意 [...$^+$] もしくは下方含意 [...$^-$]）の方向に合わせて構成素を意味的に上位もしくは下位の表現に置き換えた文と，もとの文との含意関係が成立する推論現象である．文の構成素の極性は，量化子などの論理語の性質によって定まる．例として，(1) と (2) で正解の含意関係を予測するためには，モデルは (i)*some* が上方含意の量化子であること，(ii)*every* が下方含意の量化子であること，(iii)*dog* が *animal* の下位語であること，といった，複数の意味的な性質の組み合わせを体系的に捉えている必要がある．

(1) T: ***Some*** [*dogs*$^+$] *ran yesterday*
H: ***Some*** *animals ran yesterday* **含意**

(2) T: ***Every*** [*animal*$^-$] *ran yesterday*
H: ***Every*** *dog ran yesterday* **含意**

また，関係節を含む例 (3) が示すように，言語には再帰性があり，論理的に妥当な推論のパターンは無限個考えられる．

(3) T: ***Every*** [*cat which chased* ***every*** [*dog*$^+$]] *ran*
H: ***Every*** *cat which chased* ***every*** *animal ran* **含意**

もしモデルが推論の体系性を獲得していれば，潜在的には可算無限個の推論データに対して正解を予測できるはずである．

そこで本研究では，文脈自由文法 (CFG) の生成規則からトップダウンに自動生成した推論データでモデルを学習し，(i) 未知の量化子と語彙関係の組み合わせを含む推論データと (ii) 未知の再帰構造を含む推論データに対する正答率によって，モデルの汎化性能を評価する手法を提案する．本研究の貢献は，(i)monotonicity に基づいて，自然言語の推論の体系性におけるニューラルネットの性能を評価する方法の提案，(ii) 現行のニューラル RTE モデルの，推論における汎化性能の範囲の特定，の二点である．

2 先行研究

ニューラルネットが言語の構成性を捉えられているのかという問題は古くから議論されており [9]，近年では，深層学習モデルに基づく実証的な研究が進められている．推論タスクでモデルの汎化性能を評価する先行研究では，命題論理 [11], 自然論理 [12] において，モデルが未知の単語や文長を含む推論に対して汎化性能がある可能性を示している．一方で，自然論理におけるモデルの汎化性能は限定的であり，パターンを丸暗記しているにすぎないという結果も報告されている [13]．先行研究では一条件でモデルの汎化性能を評価していたのに対して，本研究では複数条件で評価することで，モデルの汎化性能の範囲を特定する．

また，含意関係認識タスクでモデルの意味的な理解能力を評価する先行研究 [6, 7, 8, 14] では，特定の言語現象・推論現象を捉えられていない，ヒューリスティックスを用いて正解を予測している，といった，現行のモデルの問題が指摘されている．monotonicity に焦点を当ててモデルを評価する先行研究 [15, 16] もあるが，先行研究では，モデルが未知の語彙と構文の組み合わせに対してどの程度汎化するかについては着目していなかった．monotonicity は多様かつ体系的な推論パターンを扱うため，自然言語推論におけるモデルの汎化性能を評価する上で適切な推論現象である．また，下方含意の推論は結論が前提よりも長くなる傾向があり，先行研究 [14] が指摘したヒューリスティクスでは

表 1: 自動構築した推論データの例．T は CFG によって生成した文．

深さ	語彙操作	極性	引数	例
1	上位語の置換	上方含意	1	T: ***Some*** *dogs ran* H: ***Some*** *animals ran* **含意**
1	連言の追加	下方含意	2	T: ***Less than three*** *lions left* H: ***Less than three*** *lions left and cried* **含意**
2	関係節の追加	上方含意	1	T: ***Few*** *lions that hurt* ***at most three*** *dogs which ate dinner walked* H: ***Few*** *lions that hurt* ***at most three*** *dogs walked* **含意**
3	形容詞の追加	下方含意	1	T: ***Some*** *elephant* ***no*** *rabbit which touched* ***a few*** *dogs hit rushed* H: ***Some*** *elephant* ***no*** *rabbit which touched* ***a few*** *small dogs hit rushed* **含意**

解くのが困難であるため，既存のデータセットのバイアスやヒューリスティクスの問題を考慮した問題設定になっている．

3 提案手法

本研究では，学習時にモデルに見せるパターン・見せないパターンを制御してモデルの汎化性能を評価するために，CFG の生成規則から推論データを自動構築して，自動構築した推論データでモデルを学習し，評価する手法を提案する．提案手法のソースコードは研究利用が可能な形式で公開予定である．

3.1 推論データの自動構築

推論データを構築するために，まず，再帰的な規則の適用回数（以降，深さと呼ぶ）d，量化子の集合 $\mathbf{Q}$，CFG の生成規則 G によって，前提文の集合 $G_d^{\mathbf{Q}}$ を生成する．本稿では，第 1 引数と第 2 引数で上方含意の量化子 4 種類 $\mathbf{Q}^{\uparrow}$ =$\{$*some, at least three, more than three, a few*$\}$ と第 1 引数と第 2 引数で下方含意の量化子 4 種類 $\mathbf{Q}^{\downarrow}$ =$\{$*no, at most three, less than three, few*$\}$ を $\mathbf{Q}$ の要素とした．提案手法で自動構築した推論データの例を表 1 に示す．例として，$\mathbf{Q}$ の要素 *some* と生成規則 $S \rightarrow Q, N, IV$ から生成される深さ 1 の文 *Some dogs ran* は， $G_1^{\mathbf{Q}}$ の要素となる．

次に，語彙操作関数の集合 $\mathbf{L}$ の任意の要素を $G_d^{\mathbf{Q}}$ に適用することで，前提文と仮説文のペアを要素とする集合 $\mathbf{D}_d^{\mathbf{Q},\mathbf{L}}$ を生成する：

$$\mathbf{D}_d^{\mathbf{Q},\mathbf{L}} = \{(T, H) \mid T \in G_d^{\mathbf{Q}}, \exists l \in \mathbf{L}\,(l(T) = H)\}. \quad (1)$$

ここで，語彙操作とは，文中の構成素の語彙を置換または追加して意味的に上位または下位の表現にすることで仮説文を生成する操作を指す．本稿では，表 2 の 7 つの語彙操作 $\mathbf{L} = \{l_1, \ldots, l_7\}$ を対象として，各操作について語彙を 10 種類ずつ用意した．例として，前提文 *Some dogs ran* に語彙操作 l_1 を適用すると，*some* は上方含意の量化子であり，*dogs* $\sqsubseteq$ *animals* であることから，含意関係が成立する前提文と仮説文のペア *Some dogs ran* $\Rightarrow$ *Some animals ran* が生成される．このように，正解ラベルは量化子の性質と語彙関係によって自動的に決定する．加えて，本稿では一階述語論理の定理証明器 vampire[1]を用いて正解ラベルのダブルチェックを行い，証明できることを確認した．

表 2: 語彙操作の例．

関数	語彙操作	例
l_1	上位語の置換	*dogs* $\sqsubseteq$ *animals*
l_2	形容詞の追加	*small dogs* $\sqsubseteq$ *dogs*
l_3	前置詞句の追加	*dogs in the park* $\sqsubseteq$ *dogs*
l_4	関係節の追加	*dogs which ate dinner* $\sqsubseteq$ *dogs*
l_5	副詞の追加	*ran quickly* $\sqsubseteq$ *ran*
l_6	選言の追加	*ran* $\sqsubseteq$ *ran or walked*
l_7	連言の追加	*ran and barked* $\sqsubseteq$ *ran*

$\mathbf{D}_d^{\mathbf{Q},\mathbf{L}}$ のパラメータ $d, \mathbf{Q}, \mathbf{L}$ を変えて学習データ・テストデータを用意することで，モデルに見せるパターン・見せないパターンを制御できる．なお，どの評価条件においても，扱う語彙はすべて学習データとテストデータで共通にし，学習/テストデータのサイズは 300K/20K に統一する．また，学習データとテストデータにおける極性（上方含意/下方含意）の割合，正解ラベル（含意/非含意）の割合は，それぞれ 1 : 1 とする．未知の深さにおける体系性の評価においては，問題を簡単にするため，主語名詞句の関係節の埋め込みの深さの最大値を 5 として，量化子の第 1 引数の位置で語彙操作を行い構築した推論データを用いた．

3.2 未知の量化子と語彙関係の組み合わせにおける体系性

提案手法では，(i) 量化子と語彙関係の組み合わせからなる最低限の推論データでモデルを学習し評価，(ii) 量化子とすべての語彙関係の組み合わせから生成した推論データを段階的に学習データに追加してモデルを学習し評価，という二段階の評価を行うことによって，未知の量化子と語彙関係の組み合わせからなる推論におけるモデルの汎化性能の範囲を特定する．ここで最低限の推論データについて具体例で考えると，1 つの語彙関係 *dogs* $\sqsubseteq$ *animals* とすべての量化子の組み合わせからなる推論データ (4)(5)(6) と，1 つの量化子 *some* とすべての語彙関係の組み合わせからなる推論データ (4)(7) をモデルが体系的に学習していれば，未知の量化子と語彙関係の組み合わせからなる推論データ (8)(9) で正解を予測できるはずである．

[1]https://github.com/vprover/vampire

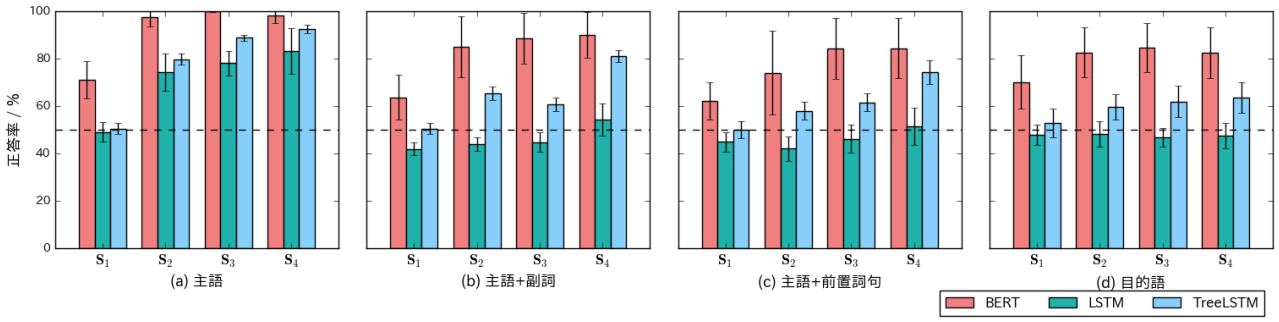

図 1: 未知の量化子と語彙関係の組み合わせにおける体系性の評価結果.

(4) T: ***Some*** *dogs ran* H: ***Some*** *animals ran* 含意
(5) T: ***Several*** *dogs ran* H: ***Several*** *animals ran* 含意
(6) T: ***No*** *animals ran* H: ***No*** *dogs ran* 含意
(7) T: ***Some*** *small dogs ran* H: ***Some*** *dogs ran* 含意
(8) T: ***No*** *dogs ran* H: ***No*** *small dogs ran* 含意
(9) T: ***Several*** *small dogs ran* H: ***Several*** *dogs ran* 含意

そこでまず， $\mathbf{Q}$ から任意の要素 q，$\mathbf{L}$ から任意の要素 l を選び，$\{q\}$ と $\mathbf{L}$ の組み合わせで生成した深さ 1 の推論データと，$\mathbf{Q}$ と $\{l\}$ の組み合わせで生成した深さ 1 の推論データの和集合 $\mathbf{D}_1^{\{q\},\mathbf{L}} \cup \mathbf{D}_1^{\mathbf{Q},\{l\}}$ を最低限の学習データ $\mathbf{S}_1$ として，$\{q\}, \{l\}$ の補集合 $\overline{\{q\}}, \overline{\{l\}}$ を用いて生成した深さ 1 の推論データの集合 $\mathbf{D}_1^{\overline{\{q\}},\overline{\{l\}}}$ でモデルを評価する．任意の q, l の $\mathbf{D}_1^{\overline{\{q\}},\overline{\{l\}}}$ に対して正解を予測できれば，モデルは未知の量化子と語彙関係の組み合わせにおける体系性を獲得している．

次に，量化子とすべての語彙関係の組み合わせから生成した推論データを段階的に学習データに追加して，類似の量化子と語彙関係の未知の組み合わせからなる推論データで正解を予測するか評価する．q を除く上方含意，下方含意の量化子 $q^{\uparrow}, q^{\downarrow}$ のペアからなる集合 $\mathbf{Q}' = \{(q^{\uparrow}, q^{\downarrow}) \mid (q^{\uparrow}, q^{\downarrow}) \subseteq \mathbf{Q}^{\uparrow} \times \mathbf{Q}^{\downarrow},\ q^{\uparrow}, q^{\downarrow} \neq q\}$ について，$\mathbf{Q}'$ の要素からなる順列の集合 $\mathsf{perm}(\mathbf{Q}')$ を考える．各 $p \in \mathsf{perm}(\mathbf{Q}')$ について，$p(i) = (q_i^{\uparrow}, q_i^{\downarrow})$ から生成した深さ 1 の推論データの集合を $2 \leq i \leq 4$ の範囲で段階的に学習データ $\mathbf{S}_i$ に追加する：

$$\mathbf{S}_{i+1} = \mathbf{S}_i \cup \mathbf{D}_1^{\{q_i^{\uparrow}, q_i^{\downarrow}\},\mathbf{L}} \quad (2 \leq i \leq 4). \qquad (2)$$

追加した学習データはテストデータから除く．

また，モデルの汎化性能が文の構成素の位置に対してどの程度頑健かについて評価するため，テストデータは (a) 例 (10) のように学習データと同じ構文構造，すなわち語彙操作が量化子を含む主語の位置で行われた場合，(b) 例 (11) のように (10) の文頭に副詞 1 語を追加した場合，(c) 例 (12) のように (10) の文頭に 3 語からなる前置詞句を追加した場合，(d) 例 (13) のように語彙操作が量化子を含む目的語の位置で行われた場合，の 4 種類の推論データを各評価条件で用意する．

(10) T: ***Several*** *small dogs ran*
H: ***Several*** *dogs ran* 含意

(11) T_{adv}: *Today* ***several*** *small dogs ran*
H_{adv}: *Today* ***several*** *dogs ran* 含意

(12) T_{prep}: *Near the shore,* ***several*** *small dogs ran*
H_{prep}: *Near the shore,* ***several*** *dogs ran* 含意

(13) T_{obj}: *Some tiger touched* ***several*** *small dogs*
H_{obj}: *Some tiger touched* ***several*** *dogs* 含意

任意の q, l で各 $|\mathsf{perm}(\mathbf{Q}')|$ 通りの実験を 5 回ずつ行い，平均正答率と標準偏差でモデルを評価する．

3.3 未知の深さにおける体系性

未知の深さの埋め込み節を含む推論における汎化性能を評価するため，$i = 1, 2$ について学習データ $\bigcup_{d \in \{1,\ldots,i+1\}} \mathbf{D}_d^{\mathbf{Q},\mathbf{L}}$ と，学習データよりも深い埋め込み節を含むテストデータ $\bigcup_{d \in \{i+2,\ldots,5\}} \mathbf{D}_d^{\mathbf{Q},\mathbf{L}}$ を用いてモデルを評価する．埋め込み節の深さ以外の要素は学習データとテストデータで共通にする．各実験を 5 回ずつ行い，平均正答率と標準偏差でモデルを評価する．

4 実験

4.1 実験設定

評価対象のモデルは LSTM [17]，TreeLSTM [4]，BERT [5] の 3 種類とした．LSTM は 3 層 LSTM を採用し，LSTM と TreeLSTM は隠れ層の次元数は 200，パラメータの更新には Adam [18]，単語ベクトルの初期値は 300 次元の学習済み GloVe [19] を使用した．BERT は pytorch の事前学習済みライブラリ[2]を用い，学習データで RTE のタスクにファインチューニングしたモデルで評価した．各モデルの学習は最大 25 エポック行い，ドロップアウトは使用しなかった．

4.2 実験結果

図 1 に未知の量化子と語彙関係の組み合わせにおける体系性の評価結果を示す．(a) をみると，最低限の学習データ $\mathbf{S}_1$ では LSTM と TreeLSTM は正答率がチャンスレート前後だったが，BERT はチャンスレートよりも向上した．さらに，学習データを追加していくと，BERT は上方含意・下方含意の量化子各 1 種

[2]https://github.com/huggingface/pytorch-pretrained-bert

表 3: 未知の深さにおける体系性の評価結果．$\mathbf{D}_d$: 再帰的な規則の適用回数 d で構築した推論データ．

学習	テスト	BERT	LSTM	TreeLSTM
$\mathbf{D}_1 \cup \mathbf{D}_2$	$\mathbf{D}_1$	100.0±0.0	100.0±0.0	100.0±0.1
	$\mathbf{D}_2$	100.0±0.0	99.8±0.2	99.5±0.1
	$\mathbf{D}_3$	75.2±10.0	75.4±10.8	86.4±4.1
	$\mathbf{D}_4$	55.0±3.7	57.7±8.7	58.6±7.8
	$\mathbf{D}_5$	49.9±4.4	45.8±4.0	48.4±3.7
$\mathbf{D}_1 \cup \mathbf{D}_2 \cup \mathbf{D}_3$	$\mathbf{D}_1$	100.0±0.0	100.0±0.0	100.0±0.0
	$\mathbf{D}_2$	100.0±0.0	95.1±7.8	99.6±0.0
	$\mathbf{D}_3$	100.0±0.0	85.2±8.9	97.7±1.1
	$\mathbf{D}_4$	77.9±10.8	59.7±10.8	68.0±5.6
	$\mathbf{D}_5$	53.5±19.6	55.1±8.2	49.6±4.3

類と全語彙関係の組み合わせから生成したデータを追加した $\mathbf{S}_2$ で正答率がほぼ 100%となり，類似の量化子について汎化することが示唆された．LSTM と TreeLSTM も学習データを追加すればするほど正答率が上がっていったが，100%には至らなかった．このことから，LSTM と TreeLSTM は，類似の量化子について完全には汎化しないことが示唆された．さらに，(b) 副詞や (c) 前置詞句を文頭に追加したテストデータや，(d) 目的語の位置で語彙操作が行われたテストデータで評価すると，BERT 含め全モデルで (a) よりも正答率が下がった．語彙は学習・テストデータで共通であり，文の構成素の位置を少し変えただけで正答率が下がったことから，すべてのモデルにおいて，量化子と語彙関係の組み合わせについて任意の構文構造のレベルでは汎化していないことが示唆された．

表 3 に未知の埋め込み節の深さにおける体系性の評価結果を示す．深さ 1, 2 を学習データとして，深さ 3 で評価した場合，すべてのモデルにおいてチャンスレートよりも正答率が上がった．しかし，深さ 4, 5 で評価した場合は，正答率がチャンスレート前後となった．同様に，深さ 1, 2, 3 を学習データとして，深さ 4 で評価した場合，すべてのモデルにおいてチャンスレートよりも正答率が上がった．しかし，深さ 5 で評価した場合は，正答率がチャンスレート前後となった．以上により，すべてのモデルにおいて，未知の深さにおける汎化性能は，学習データよりも 1 段深い埋め込みに限定されることが示唆された．また，学習データよりも 1 段深い埋め込み節における汎化性能について，モデル間で比較すると，TreeLSTM は標準偏差が小さく，汎化性能が高いことが示唆された．

5 おわりに

本研究では，ニューラルネットが自然言語の推論の体系性を学習するのかについて，monotonicity に焦点を当てて評価する手法を提案した．3 つの含意関係認識モデルを評価した結果，未知の量化子と語彙関係の組み合わせにおけるモデルの汎化性能は，学習データに含まれる文の構文構造に制限されることが示唆された．未知の深さにおけるモデルの汎化性能も，学習データよりも 1 段深い埋め込みに限定されることが示唆された．今後，多様な言語現象におけるニューラルネットの汎化性能の限界について分析を進め，任意の構文構造に対して頑健な含意関係認識モデルを検討する．

謝辞． 本研究の一部は JSPS 科研費 JP18H03284 の助成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*. Morgan & Claypool Publishers, 2013.

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

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

[4] Nam Khanh Tran and Weiwei Cheng. Multiplicative tree-structured long short-term memory networks for semantic representations. In *Proc. of *SEM*, pp. 276–286, 2018.

[5] Jacob Devlin, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL HLT*, pp. 4171–4186, 2019.

[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*, 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 *SEM*, 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.

[9] Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, Vol. 28, No. 1-2, pp. 3–71, 1988.

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

[11] Richard Evans, David Saxton, David Amos, Pushmeet Kohli, and Edward Grefenstette. Can neural networks understand logical entailment? In *Proc. of ICLR*, 2018.

[12] 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.

[13] Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. Posing fair generalization tasks for natural language inference. In *Proc. of EMNLP-IJCNLP*, pp. 4484–4494, 2019.

[14] R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proc. of ACL*, pp. 3428–3448, 2019.

[15] Richardson Kyle, Hai Hu, S. Moss Lawrence, and Sabharwal Ashish. Probing natural language inference models through semantic fragments. In *Proc. of AAAI*, 2020.

[16] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proc. of *SEM*, pp. 250–255, 2019.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

[19] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. of EMNLP*, pp. 1532–1543, 2014.