HELP: A Dataset for Identifying Shortcomings of Neural Models in Monotonicity Reasoning

Hitomi Yanaka[1,2], **Koji Mineshima**[2], **Daisuke Bekki**[2],
Kentaro Inui[1,3], **Satoshi Sekine**[1], **Lasha Abzianidze**[4], **and Johan Bos**[4]
[1]RIKEN, [2]Ochanomizu University, [3]Tohoku University, Japan
[4]University of Groningen, Netherlands
{hitomi.yanaka, 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

Abstract

Large crowdsourced datasets are widely used for training and evaluating neural models on natural language inference (NLI). Despite these efforts, neural models have a hard time capturing logical inferences, including those licensed by phrase replacements, so-called monotonicity reasoning. Since no large dataset has been developed for monotonicity reasoning, it is still unclear whether the main obstacle is the size of datasets or the model architectures themselves. To investigate this issue, we introduce a new dataset, called HELP, for handling entailments with lexical and logical phenomena. We add it to training data for the state-of-the-art neural models and evaluate them on test sets for monotonicity phenomena. The results showed that our data augmentation improved the overall accuracy. We also find that the improvement is better on monotonicity inferences with lexical replacements than on downward inferences with disjunction and modification. This suggests that some types of inferences can be improved by our data augmentation while others are immune to it.

Upward	*Some **changes in personal values** are simply part of growing older*
(MultiNLI)	⇒ *Some **changes in values** are a part of growing older*
Downward	*At most ten **commissioners** spend time at home*
(FraCaS)	⇒ *At most ten **female commissioners** spend time at home*

Table 1: Upward and downward inferences.

1 Introduction

Natural language inference (NLI) has been proposed as a benchmark task for natural language understanding. This task is to determine whether a given statement (premise) semantically entails another statement (hypothesis) (Dagan et al., 2013). Large crowdsourced datasets such as SNLI (Bowman et al., 2015a) and MultiNLI (Williams et al., 2018) have been created from naturally-occurring texts for training and testing neural models on NLI. Recent reports showed that these crowdsourced datasets contain undesired biases that allow prediction of entailment labels only from hypothesis sentences (Gururangan et al., 2018; Poliak et al., 2018b; Tsuchiya, 2018). Moreover, these standard datasets come with the so-called **upward** monotonicity inferences (see Table 1), i.e., inferences from subsets to supersets (*changes in personal values* $\sqsubseteq$ *changes in values*), but they rarely come with **downward** monotonicity inferences, i.e., inferences from supersets to subsets (*commissioners* $\sqsupseteq$ *female commissioners*). Downward monotonicity inferences are interesting in that they allow to replace a phrase with a more specific one and thus the resulting sentence can become longer, yet the inference is valid.

FraCaS (Cooper et al., 1994) contains such logically challenging problems as downward inferences. However, it is small in size (only 346 examples) for training neural models, and it covers only simple syntactic patterns with severely restricted vocabularies. The lack of such a dataset on a large scale is due to at least two factors: it is hard to instruct crowd workers without deep knowledge of natural language syntax and semantics, and it is also unfeasible to employ experts to obtain a large number of logically challenging inferences.

Bowman et al. (2015b) proposed an artificial dataset for logical reasoning, whose premise and hypothesis are automatically generated from a simple English-like grammar. Following this line of work, Geiger et al. (2018) presented a method to construct a complex dataset for multiple quantifiers (e.g., *Every dwarf licks no rifle* ⇒ *No ugly dwarf licks some rifle*). These datasets contain downward inferences, but they are designed not to require lexical knowledge. There are also NLI datasets which expand lexical knowledge by replacing words using lexical rules (Monz and de Rijke, 2001; Glockner et al., 2018; Naik et al., 2018;

*Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 250–255
Minneapolis, June 6–7, 2019. ©2019 Association for Computational Linguistics

Poliak et al., 2018a). In these works, however, little attention has been paid to downward inferences.

The GLUE leaderboard (Wang et al., 2019) reported that neural models did not perform well on downward inferences, and this leaves us guessing whether the lack of large datasets for such kind of inferences that involve the interaction between lexical and logical inferences is an obstacle of understanding inferences for neural models.

To shed light on this problem, this paper makes the following three contributions: (a) providing a method to create a large NLI dataset[1] that embodies the combination of lexical and logical inferences focusing on monotonicity (i.e., phrase replacement-based reasoning) (Section 3), (b) measuring to what extent the new dataset helps neural models to learn monotonicity inferences, and (c) by analyzing the results, revealing which types of logical inferences are solved with our training data augmentation and which ones are immune to it (Section 4.2).

2 Monotonicity Reasoning

Monotonicity reasoning is a sort of reasoning based on word replacement. Based on the monotonicity properties of words, it determines whether a certain word replacement results in a sentence entailed from the original one (van Benthem, 1983; Icard and Moss, 2014). A polarity is a characteristic of a word position imposed by monotone operators. Replacements with more general (or specific) phrases in ↑ (or ↓) polarity positions license entailment. Polarities are determined by a function which is always upward monotone $(+)$ (i.e., an order preserving function that licenses entailment from specific to general phrases), always downward monotone $(-)$ (i.e., an order reversing function) or neither, non-monotone.

Determiners are modeled as binary operators, taking noun and verb phrases as the first and second arguments, respectively, and they entail sentences with their arguments narrowed or broadened according to their monotonicity properties. For example, the determiner *some* is upward monotone both in its first and second arguments, and the concepts can be broadened by replacing its hypernym (*people* $\sqsupseteq$ *boy*), removing modifiers (*dancing* $\sqsupseteq$ *happily dancing*), or adding disjunction. The concepts can be narrowed by replacing its hyponym (*schoolboy* $\sqsubseteq$ *boy*), adding modifiers, or adding conjunction.

(1) *Some* $[_{\text{NP}}$ *boys*$\uparrow]^{+}[_{\text{VP}}$ *are happily dancing*$\uparrow]^{+}$
$\Rightarrow$ *Some* $[_{\text{NP}}$ *people*$]$ $[_{\text{VP}}$ *are dancing*$]$
$\not\Rightarrow$ *Some* $[_{\text{NP}}$ *schoolboys*$]$ $[_{\text{VP}}$ *are dancing and singing*$]$

If a sentence contains negation, the polarity of words over the scope of negation is reversed:

(2) *No* $[_{\text{NP}}$ *boys*$\downarrow]^{-}[_{\text{VP}}$ *are happily dancing*$\downarrow]^{-}$
$\not\Rightarrow$ *No* $[_{\text{NP}}$ *one*$]$ $[_{\text{VP}}$ *is dancing*$]$
$\Rightarrow$ *No* $[_{\text{NP}}$ *schoolboys*$]$ $[_{\text{VP}}$ *are dancing and singing*$]$

If the propositional object is embedded in another negative or conditional context, the polarity of words over its scope can be reversed again:

(3) *If* [*there are no* $[_{\text{NP}}$ *boys*$\uparrow]^{-}[_{\text{VP}}$ *dancing happily*$\uparrow]^{-}]^{-}$,
[*the party might be canceled*]$^{+}$
$\Rightarrow$ *If* [*there is no* $[_{\text{NP}}$ *one*$]$ $[_{\text{VP}}$ *dancing*$]]$,
[*the party might be canceled*]

In this way, the polarity of words is determined by monotonicity operators and syntactic structures.

3 Data Creation

We address three issues when creating the inference problems: (a) Detect the monotone operators and their arguments; (b) Based on the syntactic structure, induce the polarity of the argument positions; (c) Using lexical knowledge or logical connectives, narrow or broaden the arguments.

3.1 Source corpus

We use sentences from the Parallel Meaning Bank (PMB, Abzianidze et al., 2017) as a source while creating the inference dataset. The reason behind choosing the PMB is threefold. First, the fine-grained annotations in the PMB facilitate our automatic monotonicity-driven construction of inference problems. In particular, semantic tokenization and WordNet (Fellbaum, 1998) senses make narrow and broad concept substitutions easy while the syntactic analyses in Combinatory Categorial Grammar (CCG, Steedman, 2000) format and semantic tags (Abzianidze and Bos, 2017) contribute to monotonicity and polarity detection. Second, the PMB contains lexically and syntactically diverse texts from a wide range of genres. Third, the gold (silver) documents are fully (partially) manually verified, which control noise in the automated generated dataset. To prevent easy inferences, we use the sentences with more than five tokens from 5K gold and 5K silver portions of the PMB.

[1] Our dataset and its generation code will be made publicly available at https://github.com/verypluming/HELP.

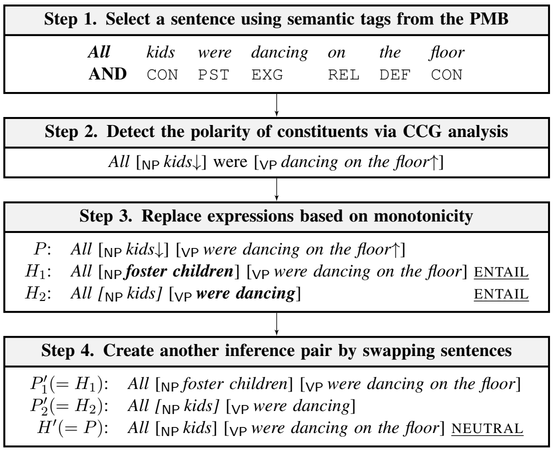
Step 1. Select a sentence using semantic tags from the PMB
All kids were dancing on the floor
AND CON PST EXG REL DEF CON
Step 2. Detect the polarity of constituents via CCG analysis
All [NP kids↓] were [VP dancing on the floor↑]
Step 3. Replace expressions based on monotonicity
P: All [NP kids↓] [VP were dancing on the floor↑]
H1: All [NP foster children] [VP were dancing on the floor] ENTAIL
H2: All [NP kids] [VP were dancing] ENTAIL
Step 4. Create another inference pair by swapping sentences
P′1(= H1): All [NP foster children] [VP were dancing on the floor]
P′2(= H2): All [NP kids] [VP were dancing]
H′(= P): All [NP kids] [VP were dancing on the floor] NEUTRAL

Figure 1: Illustration for creating the HELP dataset.

3.2 Methodology

Figure 1 illustrates the method of creating the HELP dataset. We use declarative sentences from the PMB containing monotone operators, conjunction, or disjunction as a source (Step 1). These target words can be identified by their semantic tags: AND (*all*, *every*, *each*, *and*), DIS (*some*, *several*, *or*), NEG (*no*, *not*, *neither*, *without*), DEF (*both*), QUV (*many*, *few*), and IMP (*if*, *when*, *unless*). In Step 2, after locating the first (NP) and the second (VP) arguments of the monotone operator via a CCG derivation, we detect their polarities with the possibility of reversing a polarity if an argument appears in a downward environment.

In Step 3, to broaden or narrow the first and the second arguments, we consider two types of operations: (i) lexical replacement, i.e., substituting the argument with its hypernym/hyponym (e.g., H_1) and (ii) syntactic elimination, i.e., dropping a modifier or a conjunction/disjunction phrase in the argument (e.g., H_2). Given the polarity of the argument position (↑ or ↓) and the type of replacement (with more general or specific phrases), the gold label (*entailment* or *neutral*) of a premise-hypothesis pair is automatically determined; e.g., both (P, H_1) and (P, H_2) in Step 3 are assigned *entailment*. For lexical replacements, we use WordNet senses from the PMB and their ISA relations with the same part-of-speech to control naturalness of the obtained sentence. To compensate missing word senses from the silver documents, we use the Lesk algorithm (Lesk, 1986). In Step 4, by swapping the premise and the hypothesis, we create another inference pair and assign its gold label; e.g., (P'_1, H') and (P'_2, H') are created and assigned *neutral*. By swapping a sentence pair created by syntactic elimination, we can create a pair such as (P'_2, H') in which the hypothesis is more specific than the premise.

Section	Size	Example
Up	7784	*Tom bought some **Mexican sunflowers** for Mary* ⇒*Tom bought some **flowers** for Mary**
Down	21192	*If there's no **water**, there's no whisky** ⇒*If there's no **facility**, there's no whisky*
Non	1105	*Shakespeare wrote both **tragedy and comedy**** ⇏*Shakespeare wrote both **tragedy and drama***
Conj	6076	*Tom removed his glasses* ⇏*Tom removed his glasses **and rubbed his eyes****
Disj	438	*The trees are barren* ⇒*The trees are barren **or bear only small fruit****

Table 2: Examples in HELP. The sentence with an asterisk is the original sentence from the PMB.

3.3 The HELP dataset

The resulting dataset has 36K inference pairs consisting of upward monotone, downward monotone, non-monotone, conjunction, and disjunction. Table 2 shows some examples. The number of vocabulary items is 15K. We manually checked the naturalness of randomly sampled 500 sentence pairs, of which 146 pairs were unnatural. As mentioned in previous work (Glockner et al., 2018), there are some cases where WordNet for substitution leads to unnatural sentences due to the context mismatch; e.g., an example such as *P*: *You have no driving happening* ⇒ *H*: *You have no driving experience*, where *P* is obtained from *H* by replacing *experience* by its hypernym *happening*. Since our intention is to explore possible ways to augment training data for monotonicity reasoning, we include these cases in the training dataset.

4 Experiments

We use HELP as additional training material for three neural models for NLI and evaluate them on test sets dealing with monotonicity reasoning.

4.1 Experimental settings

Models We used three models: BERT (Devlin et al., 2019), BiLSTM+ELMo+Attn (Wang et al., 2019), and ESIM (Chen et al., 2017).

Training data We used three different training sets and compared their performance; MultiNLI (392K), MultiNLI+MQ (the dataset for multiple quantifiers introduced in Section 1; Geiger et al., 2018) (892K), and MultiNLI+HELP (429K).

Test data We used four test sets: (i) the GLUE diagnostic dataset (Wang et al., 2019) (upward monotone, downward monotone, non-monotone,

Model	Train Data	GLUE diagnostic												FraCaS		SICK		MNLI			
		Up (30)		Down (30)		Non (22)		Conj (32)		Disj (38)		Total (152)		(80)		(4927)		match (10000)		mismatch (10000)	
			△		△		△		△		△		△		△		△		△		△
	MNLI	50.4		-67.5		23.1		52.5		-6.1		17.8		65.0		55.4		**84.6**		**83.4**	
BERT	+MQ	59.6	+9.2	-49.3	+18.2	14.0	-9.1	62.1	+9.6	-18.8	-12.7	26.3	+8.5	**68.8**	+3.8	58.2	+2.8	78.4	-6.2	78.6	-4.8
	+HELP	**67.0**	+16.6	**29.8**	+97.3	**47.9**	+24.8	**72.1**	+19.6	**-4.1**	+2.0	**51.2**	+33.4	**68.8**	+3.8	**60.0**	+4.6	84.4	-0.2	83.1	-0.3
BiLSTM	MNLI	22.2		-9.4		-2.7		42.4		-9.9		-3.5		68.9		53.8		**76.4**		**76.1**	
+ELMo	+MQ	22.2	0.0	8.1	+17.5	-5.7	-3.0	42.4	0.0	**-9.8**	+0.1	5.7	+9.2	65.9	-3.0	**54.0**	+0.2	71.4	-5.0	70.7	-5.4
+Attn	+HELP	**32.4**	+10.2	**22.9**	+32.3	**3.7**	+6.4	**45.6**	+3.2	-9.9	0.0	**17.0**	+20.5	**71.3**	+2.4	**54.0**	+0.2	75.2	-1.2	74.1	-2.0
	MNLI	14.9		-14.0		6.0		29.8		-3.6		1.1		47.5		43.9		**71.3**		**70.7**	
ESIM	+MQ	27.2	+12.3	-7.8	+6.2	3.4	-2.6	5.2	-24.6	-13.0	-9.4	6.8	+5.7	43.7	-3.8	53.1	+9.2	68.6	-3.7	68.2	-2.5
	+HELP	**31.4**	+16.5	**24.7**	+38.7	**8.0**	+2.0	**32.6**	+2.8	**7.1**	+10.7	**27.0**	+25.9	**48.8**	+1.3	**56.6**	+12.7	71.1	-0.2	70.1	-0.6

Table 3: Evaluation results on the GLUE diagnostic dataset, FraCaS, SICK, and MultiNLI (MNLI). The number in parentheses is the number of problems in each test set. △ is the difference from the model trained on MNLI.

conjunction, and disjunction sections), (ii) FraCaS (the generalized quantifier section), (iii) the SICK (Marelli et al., 2014) test set, and (iv) MultiNLI matched/mismatched test set. We used the Matthews correlation coefficient (ranging $[-1, 1]$) as the evaluation metric for GLUE. Regarding other datasets, we used accuracy as the metric. We also check if our data augmentation does not decrease the performance on MultiNLI.

4.2 Results and discussion

Table 3 shows that adding HELP to MultiNLI improved the accuracy of all models on GLUE, FraCaS, and SICK. Regarding MultiNLI, note that adding data for downward inference can be harmful for performing upward inference, because lexical replacements work in an opposite way in downward environments. However, our data augmentation minimized the decrease in performance on MultiNLI. This suggests that models managed to learn the relationships between downward operators and their arguments from HELP.

The improvement in accuracy is better with HELP than that with MQ despite the fact that the size of HELP is much smaller than MQ. MQ does not deal with lexical replacements, and thus the improvement is not stable. This indicates that the improvement comes from carefully controlling the target reasoning of the training set rather than from its size. ESIM showed a greater improvement in accuracy compared with the other models when we added HELP. This result arguably supports the finding in Bowman et al. (2015b) that a tree architecture is better for learning some logical inferences. Regarding the evaluation on SICK, Talman and Chatzikyriakidis (2018) reported a drop in accuracy of 40-50% when BiLSTM and ESIM were trained on MultiNLI because SICK is out of the domain of MultiNLI. Indeed, the accuracy of each model, including BERT, was low at 40-60%.

When compared among linguistic phenomena, the improvement by adding HELP was better for upward and downward monotone. In particular, all models except models trained with HELP failed to answer 68 problems for monotonicity inferences with lexical replacements. This indicates that such inferences can be improved by adding HELP.

The improvement for disjunction was smaller than other phenomena. To investigate this, we conducted error analysis on 68 problems of GLUE and FraCaS, which all the models misclassified. 44 problems are neutral problems in which all words in the hypothesis occur in the premise (e.g., *He is either in London or in Paris* $\not\Rightarrow$ *He is in London*). 13 problems are entailment problems in which the hypothesis contains a word or a phrase not occurring in the premise (e.g., *I don't want to have to keep entertaining people* $\Rightarrow$ *I don't want to have to keep entertaining people who don't value my time*). These problems contain disjunction or modifiers in downward environments where either (i) the premise P contains all words in the hypothesis H yet the inference is invalid or (ii) H contains more words than those in P yet the inference is valid.[2] Although HELP contains 21K such problems, the models nevertheless misclassified them. This indicates that the difficulty in learning these non-lexical downward inferences might not come from the lack of training datasets.

5 Conclusion and Future Work

We introduced a monotonicity-driven NLI data augmentation method. The experiments showed that neural models trained on HELP obtained the higher overall accuracy. However, the improvement tended to be small on downward monotone inferences with disjunction and modification, which suggests that some types of inferences can

[2] Interestingly, certain logical inferences including disjunction and downward monotonicity are difficult also for humans to get (Geurts and van der Slik, 2005).

be improved by adding data while others might require different kind of *help*.

For future work, our data augmentation can be used for multilingual corpora. Since the PMB annotations sufficed for creating HELP, applying our method to the non-English PMB documents seems straightforward. Additionally, it is interesting to verify the quality and contribution of a dataset which will be created by using our method on an automatically annotated and parsed corpus.

Acknowledgement

We thank our three anonymous reviewers for helpful suggestions.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 242–247.

Lasha Abzianidze and Johan Bos. 2017. Towards universal semantic tagging. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*, pages 1–6.

Johan van Benthem. 1983. Determiners and logic. *Linguistics and Philosophy*, 6(4):447–478.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

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

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1657–1668.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

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

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2018. Stress-testing neural models of natural language inference with multiply-quantified sentences. *CoRR*, abs/1810.13033.

Bart Geurts and Frans van der Slik. 2005. Monotonicity and processing load. *Journal of Semantics*, 22(1):97–117.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 650–655.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Thomas Icard and Lawrence Moss. 2014. Recent progress in monotonicity. *LILT*, 9(7):167–194.

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

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 216–223.

Christof Monz and Maarten de Rijke. 2001. Lightweight entailment checking for computational semantics. In *Proceedings of the 3rd International Workshop on Inference in Computational Semantics*, pages 1–15.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 337–340.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Aarne Talman and Stergios Chatzikyriakidis. 2018. Testing the generalization power of neural network models across NLI benchmarks. *CoRR*, abs/1810.09774.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1112–1122.