Do Neural Models Learn Transitivity of Veridical Inference?

Hitomi Yanaka[1,2], **Koji Mineshima**[3], **and Kentaro Inui**[4,2]
[1]The University of Tokyo, [2]RIKEN, [3]Keio University, [4]Tohoku University
hyanaka@is.s.u-tokyo.ac.jp, minesima@abelard.flet.keio.ac.jp,
inui@ecei.tohoku.ac.jp

Introduction Central to human-like generalization capacities is the fact that ability to understand a sentence is related to ability to understand other sentences, called *systematicity* of human cognition in Fodor and Pylyshyn (1988). We explore whether DNN models possess this type of generalization capacity in the domain of natural language inference (NLI), which is the task to judge whether a premise entails a hypothesis (Dagan et al., 2013; Bowman et al., 2015). A key property underlying systematicity of drawing inferences is the *transitivity* of inference relations, illustrated in Figure 1. Schematically, if a model learns a basic inference pattern from A to B and one from B to C, it should be able to compose the two patterns to draw a new inference from A to C. If a model lacks this generalization capacity, it must memorize an exponential number of inference combinations independently of basic patterns.

We focus on transitivity inferences that combine *veridical* inferences with other types. In veridical inferences, one must distinguish two entailment types. For example, the verb *know* is called **veridical** in that "x *knows that* P" entails that P is true, while the verb *hope* is called **non-veridical** since "x *hopes that* P" does not entail that P is true. Veridical inferences can relatively easily compose transitivity inferences at scale by embedding various inference types into clause-embedding verbs. As Figure 1 shows, if a model has the ability to perform both Boolean inference and veridical inference, it is desirable to have the ability to combine both types to make a chained inference. Such transitivity inferences are by no means trivial. If the premise is changed to *Jo knows that Ann* ***or*** *Bob left*, it does not follow that *Bob left*, even though the veridical verb *know* appears. Models relying on shallow heuristics such as lexical overlap can wrongly predict *entailment* in this case. To correctly handle such composite inferences, models must capture structural relations between veridical inferences and various kinds of embedded inference.

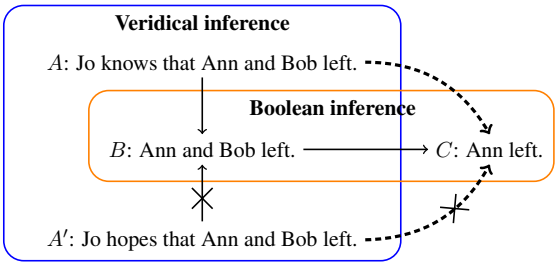

Figure 1: Illustration of transitivity inferences (indicated by ⇢) composed of two basic inferences, veridical and Boolean. Arrows indicate *entailment* and arrows with a cross (×) indicate *non-entailment*.

We create and publicly release two types of NLI datasets for testing model ability to perform transitivity inferences: a fully synthetic dataset that combines veridical inferences and Boolean inferences, and a naturalistic dataset that combines veridical inferences with lexical and structural inferences. We use these datasets to analyze whether standard NLI models can perform transitivity inference.

Overview We consider two basic inference patterns and their combinations. The first basic pattern, $\mathcal{I}_1$, is veridical inference. We write $f(s_1) \rightarrow s_1$ to denote a schematic veridical inference, where f is a clause-embedding verb and s_1 is the embedded clause. For instance, in the case of the inference pattern $A \rightarrow B$ in Figure 1, "*Jo knows that x*" corresponds to $f(x)$ and "*Ann and Bob left*" to s_1. The second basic pattern, $\mathcal{I}_2$, provides an inference from the embedded material. We denote a premise-hypothesis pair of this second inference by $s_1 \rightarrow s_2$.

Given two inferences $f(s_1) \rightarrow s_1$ in $\mathcal{I}_1$ and $s_1 \rightarrow s_2$ in $\mathcal{I}_2$, we consider a new inference $f(s_1) \rightarrow s_2$, where premise $f(s_1)$ is the same as that of $\mathcal{I}_1$ and

$f(s_1) \rightarrow s_1$	$s_1 \rightarrow s_2$	$f(s_1) \rightarrow s_2$	Example
yes	*yes*	*yes*	$f(s_1)$: Someone **realized** that [a boy was playing a guitar]. s_1: A boy was playing a guitar. s_2: A kid was playing a guitar.
unk	*yes*	*unk*	$f(s_1)$: Someone **doubts** that [the woman is putting makeup on the man]. s_1: The woman is putting makeup on the man. s_2: A man's face is being painted by a woman.
yes	*unk*	*unk*	$f(s_1)$: Someone **remembered** that [a cat was playing with a device]. s_1: A cat was playing with a device. s_2: The boy was enthusiastically playing in the mud.

Table 1: Examples of our transitivity inference set.

Type			Model	
$f(s_1) \rightarrow s_1$	$s_1 \rightarrow s_2$	$f(s_1) \rightarrow s_2$	**LSTM**	**BERT**
yes	*yes*	***yes***	89.0 ± 9.1	100.0 ± 0.0
yes	*unk*	***unk***	6.3 ± 12.8	0.4 ± 7.8
unk	*yes*	***unk***	93.4 ± 8.3	99.4 ± 9.0
unk	*unk*	***unk***	92.1 ± 7.2	99.5 ± 0.5
Total			70.2 ± 3.4	84.2 ± 1.2
Validation ($f(s_1) \rightarrow s_1$)			82.1 ± 3.3	99.2 ± 0.0
Validation ($s_1 \rightarrow s_2$)			81.9 ± 3.0	89.1 ± 2.0

Table 2: Results on the synthetic transitivity test set.

Type			Model		Human
$f(s_1) \rightarrow s_1$	$s_1 \rightarrow s_2$	$f(s_1) \rightarrow s_2$	**LSTM**	**BERT**	
yes	*yes*	***yes***	97.1 ± 2.7	100.0 ± 0.0	98.8
yes	*unk*	***unk***	0.0 ± 0.0	8.9 ± 7.8	98.8
unk	*yes*	***unk***	97.1 ± 2.7	100.0 ± 0.0	44.9
unk	*unk*	***unk***	97.3 ± 2.6	100.0 ± 0.0	99.6
Total			73.5 ± 1.6	75.2 ± 1.7	85.5
Validation ($f(s_1) \rightarrow s_1$)			82.1 ± 3.3	99.2 ± 0.0	
Validation ($s_1 \rightarrow s_2$)			81.9 ± 3.0	89.1 ± 2.0	

Table 3: Results on the naturalistic transitivity test set.

hypothesis s_2 is the same as that of $\mathcal{I}_2$. See Table 1 for inference examples $f(s_1) \rightarrow s_1$, $s_1 \rightarrow s_2$, and $f(s_1) \rightarrow s_2$. We consider binary labels, *entailment* and *non-entailment*, denoted by *yes* and *unk*, respectively. As Table 1 shows, the gold label on the $f(s_1) \rightarrow s_2$ pattern can be determined from those of the basic patterns $f(s_1) \rightarrow s_1$ and $s_1 \rightarrow s_2$, following the transitivity of entailment relations.

We train models with the first and second patterns, $f(s_1) \rightarrow s_1$ and $s_1 \rightarrow s_2$, and then test them on a set of the composite inferences $f(s_1) \rightarrow s_2$ that combines them. Model capable of applying the transitivity inference from $f(s_1) \rightarrow s_1$ and $s_1 \rightarrow s_2$ to $f(s_1) \rightarrow s_2$ should consistently predict the correct label of $f(s_1) \rightarrow s_2$ for any combination of $f(s_1) \rightarrow s_1$ and $s_1 \rightarrow s_2$.

Datasets We collect 30 clause-embedding verbs f that take tensed subordinate clauses appearing in both MegaVeridicality2 (White et al., 2018) and the verb veridicality dataset (Ross and Pavlick, 2019). To test diverse inference patterns, we consider two types of the second basic inference $s_1 \rightarrow s_2$: synthesized Boolean inferences and naturalistic inferences using an existing NLI dataset, SICK (Marelli et al., 2014), which covers various lexical and structural inference. As shown in examples in Table 1, we can generate a new sentence $f(s_1)$ by selecting a clause-embedding verb f and a premise sentence s_1 of the second basic inference. Then, we can obtain a veridical inference example $f(s_1) \rightarrow s_1$ by setting $f(s_1)$ as a premise and s_1 as a hypothesis. Likewise, we can obtain a composite inference example $f(s_1) \rightarrow s_2$. We provide 6,000 $f(s_1) \rightarrow s_2$ examples for fully synthetic datasets and 30,000 $f(s_1) \rightarrow s_2$ examples for naturalistic datasets. The ratio of the gold labels (*yes* and *unk*) of the training set containing $f(s_1) \rightarrow s_1$ and $s_1 \rightarrow s_2$ examples is $1:1$, and the ratio for the $f(s_1) \rightarrow s_2$ test set is $1:3$.

Experiments and Analysis We analyze whether two standard NLI models (LSTM and BERT) trained with the basic inference set can consistently perform composite inferences on the test set. Table 2 and Table 3 shows that the models trained with the basic inference set performed substantially below chance for the examples $f(s_1) \rightarrow s_2$ where f is veridical and $s_1 \rightarrow s_2$ is *unk*. This suggests that while the models achieve over 80% accuracy on both $f(s_1) \rightarrow s_1$ and $s_1 \rightarrow s_2$ validation sets, they do not apply transitivity inference from the inferences $f(s_1) \rightarrow s_1$ and $s_1 \rightarrow s_2$, but rather predict the label for the composite inference $f(s_1) \rightarrow s_2$ by judging whether it is similar to the veridical inference $f(s_1) \rightarrow s_1$ in the training set. We also collect human judgements for our naturalistic transitivity test set, and human performance is near perfect for the examples $f(s_1) \rightarrow s_2$ where f is veridical and $s_1 \rightarrow s_2$ is *unk*.

Conclusion We introduced an analysis method using transitivity inferences for evaluating the systematic generalization capacities of NLI models. Experiments showed that current NLI models fail to consistently perform transitive inference. This indicates that there is still much room for improving systematic generalization capacities of NLI models with respect to combining basic inferential abilities on various linguistic phenomena.

Acknowledgement We thank the three anonymous reviewers for their helpful comments and suggestions. This work was partially supported by JSPS KAKENHI Grant Number JP20K19868.

References

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

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.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

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 Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.