Does BERT Know that the IS-A Relation Is Transitive?

Ruixi Lin and **Hwee Tou Ng**
Department of Computer Science
National University of Singapore
{ruixi,nght}@comp.nus.edu.sg

Abstract

The success of a natural language processing (NLP) system on a task does not amount to fully understanding the complexity of the task, typified by many deep learning models. One such question is: can a black-box model make logically consistent predictions for transitive relations? Recent studies suggest that pre-trained BERT can capture lexico-semantic clues from words in the context. However, to what extent BERT captures the transitive nature of some lexical relations is unclear. From a probing perspective, we examine WordNet word senses and the IS-A relation, which is a transitive relation. That is, for senses A, B, and C, A *is-a* B and B *is-a* C entail A *is-a* C. We aim to quantify how much BERT agrees with the transitive property of IS-A relations, via a minimalist probing setting. Our investigation reveals that BERT's predictions do not fully obey the transitivity property of the IS-A relation.[1]

1 Introduction

The IS-A relation denotes a subclass relation. If A *is-a* B, then the concept A is a subclass of the concept B, or A is subsumed by B. The IS-A relation is frequently encoded in lexical taxonomies. The IS-A relation has great significance since it empowers generalization, and generalization is at the core of machine inference for text understanding. The IS-A hierarchy is inherently transitive, i.e., for three concepts (or word senses) A, B, and C, A *is-a* B and B *is-a* C entail A *is-a* C. For example, knowing that *humanoid* is a type of *automaton*, and *automaton* is a type of *artifact*, then by transitivity, the relation *humanoid* is an *artifact* also holds.

The concept of transitivity is easy to comprehend by humans. However, deep learning models, including pre-trained language models such as BERT (Devlin et al., 2019), are known to lack some human-level generalization capacities in text understanding, or it may show some capacities for making correct predictions but for the wrong reasons, including being insensitive to negation and exploiting only surface features (Kassner and Schutze, 2020; Ettinger, 2020), lacking understanding of perceptual properties (Forbes et al., 2019; Weir et al., 2020), and surface form competition (Holtzman et al., 2021).

Despite the issues raised above, previous work has shown that BERT's layers align with the NLP pipeline, and representations in the different layers of BERT are found to capture different levels of textual understanding, from syntactic (e.g., part-of-speech tagging) to semantic (e.g., semantic role labeling) as the layers go from the lower to higher layers (Tenney et al., 2019a,b). Recent studies also suggest that BERT can capture lexical relation clues from words in contexts (Vulić et al., 2020; Misra et al., 2020). Researchers begin to recognize BERT as an open knowledge source and query BERT for information (Petroni et al., 2019). Moreover, BERT, even without fine-tuning on downstream tasks, possesses a fair ability to produce contextualized embeddings that cluster to word senses (Wiedemann et al., 2019; Haber and Poesio, 2020; Mickus et al., 2020; Loureiro et al., 2021). These findings suggest that BERT has some understanding of the building blocks of language. Following these findings, since an IS-A taxonomy can be built on top of explicit word senses, do contextualized embeddings learned from BERT for word senses (in particular contexts) respect the properties of the IS-A taxonomy, specifically transitivity? That is, does BERT make logically consistent predictions that enforce the transitivity constraint of the IS-A relation?

In this paper, we introduce a minimalist probing method to investigate whether BERT knows that the IS-A relation is transitive. We first quantify how well BERT predicts the IS-A relation. Next,

[1] The source code and dataset of this paper are available at https://github.com/nusnlp/probe-bert-transitivity.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 94 - 99
May 22-27, 2022 ©2022 Association for Computational Linguistics

we measure the extent to which BERT enforces the transitivity constraint. That is, given that BERT predicts A *is-a* B and B *is-a* C, does it then predict A *is-a* C?

In our work, we make use of WordNet (Fellbaum, 1998) and propose a method to sample word sense pairs with contexts from WordNet example sentences to build a probing dataset. We use a nearest neighbor classifier for probing, which does not require any parameter tuning. Our findings indicate that BERT can predict IS-A relations with an accuracy score of 72.6%. However, when BERT predicts A *is-a* B and B *is-a* C, it only predicts A *is-a* C 82.4% of the time. This suggests that simply treating BERT as is as a knowledge base (Petroni et al., 2019) is not completely satisfactory, and additional work needs to be done to incorporate the transitivity constraint in natural language inference when using BERT.

2 Related Work

A key weakness of deep learning models is that they are black-box models and do not offer explainable and interpretable predictions. This has led to a large body of research regarding their interpretability (Linardatos et al., 2021). The pre-trained language model BERT has been extensively analyzed since its release. In particular, feature-based probes have been proposed to show how a particular layer, head, or neuron of BERT works on a downstream NLP task. Usually with a small set of additional parameters, a probe is trained in a supervised manner using feature representations from the pre-trained BERT, e.g., contextualized embeddings, to solve a particular task (Wu et al., 2020). Attention and structural probes have been invented to investigate different aspects of BERT and linguistic properties (Lin et al., 2019; Jawahar et al., 2019; Clark et al., 2019; Hewitt and Manning, 2019; Manning et al., 2020; Tenney et al., 2019a,b; Pruksachatkun et al., 2020). Latent ontology of contextual embeddings has also been investigated via cluster analysis (Michael et al., 2020). On probing contextualized representations for lexico-semantic relations, previous studies have investigated BERT for lexical relation classification via a neural network probe on type-level embeddings (Vulić et al., 2020).

Our work differs from prior work by our goal to explicitly investigate how much BERT understands the IS-A relation and more importantly, obeys the transitivity constraint. That is, we aim to determine if and how often BERT makes logically consistent predictions for the IS-A relation. Moreover, we focus on investigating sense-based IS-A relation, which is associated with explicit word senses in their contexts, so contextualized embeddings can be clearly mapped to word senses.

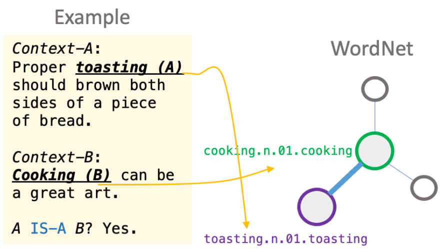
Example
Context-A:
Proper toasting (A) should brown both sides of a piece of bread.
Context-B:
Cooking (B) can be a great art.
A IS-A B? Yes.
WordNet
cooking.n.01.cooking
toasting.n.01.toasting

Figure 1: An example of an IS-A pair

3 Experimental Setup

We probe if and how well BERT can predict the IS-A relation and its transitivity.

Task Definition For a dataset of interest, we denote it as $\mathcal{D} = \{[(u_1, v_1), y_1], \cdots, [(u_n, v_n), y_n]\}$. $(u_{1:n}, v_{1:n}) = [(u_1, v_1) \cdots, (u_n, v_n)]$ are representations for pairs of word senses and $y_{1:n} = (y_1, \cdots, y_n)$ are labels for the IS-A relation classification task, where 1 denotes the positive IS-A relation and 0 otherwise. In our probing task, we quantify the extent to which $(u_{1:n}, v_{1:n})$ encode relations $y_{1:n}$. To probe BERT, we use the contextualized embedding (i.e., BERT's final hidden state output) of a word in a given context as the representation for the sense associated with the word's meaning in that context.

Contextualized Embeddings In this work, we focus on the BERT-base model. Given a target word w_t and its context c, BERT produces a final hidden state output as the contextualized embedding o_t for the target word w_t. If w_t is tokenized into subwords, we take the average over all subwords to be the contextualized embedding.

3.1 Probing Dataset

WordNet (Fellbaum, 1998) is a rich lexical database of word senses connected via the IS-A relation with example sentences for sense usages, making it a natural resource for probing. A 1-hop IS-A relation is illustrated in Figure 1. By transitivity, an n-hop IS-A relation is formed from a chain of n parent-child IS-A links. We focus on noun pairs in this work as nouns make up 70% of all senses in WordNet, and the path lengths for nouns are often longer than for verbs. We propose a path-based sampling method to generate pairs from WordNet, as follows:

1. Let $\mathcal{L} = \{ s \mid s \in \mathcal{S}$ and hypo(s) = $\emptyset \}$ denote all leaf senses from WordNet, where $\mathcal{S}$ represents the set of all senses in WordNet, and hypo(s) denotes the set of hyponym (children) senses of sense s.

2. For IS-A, sample a leaf sense N uniformly at random from $\mathcal{L}$, connect N to the root R, which gives a path p ($N \rightarrow R$). For not IS-A, similarly sample two leaf senses N_1, N_2 randomly from $\mathcal{L}$ and obtain two paths p_1 ($N_1 \rightarrow R$) and p_2 ($N_2 \rightarrow R$).

3. For IS-A, randomly sample three senses A, B, C from p and ensure that example sentences exist for senses A, B, C. This results in the 3-tuple (A, B, C) and three positive examples $(A, B), (B, C), (A, C)$. For not IS-A, randomly sample A' and B' from p_1 and p_2 respectively, ensuring that example sentences exist for senses A' and B'. If A' is not on the path of $B' \rightarrow R$ and vice versa, then we obtain a negative example (A', B'); else return to step 2.

4. Repeat step 2 and 3 to sample more positive and negative examples, until the desired number of examples is reached.

In our probing dataset, We have 1,665 3-tuples resulting in 4,995 positive examples, as well as 4,995 negative examples, where each example is a pair of senses.

3.2 Probing Method

Since our goal is to determine what BERT as a pretrained language model knows about transitivity, we use a simple nearest neighbor (1-nn) classifier without further fine-tuning of BERT's parameters. We also adopt a 1-nn classifier instead of a more complex classifier so that we are measuring what BERT knows and not what is learned by a subsequent complex classification model.

Our 1-nn probing classifier works by finding the closest example in the training set for a test example, and using the closest training example's label as the prediction. Euclidean distance is used as the distance metric for our 1-nn probing classifier. We represent each example, which is a pair of senses, by the concatenation of the contextualized embeddings of the pair. For a pair of target words (w_1, w_2) and their respective contextualized embeddings (o_1, o_2), $r(o_1, o_2)$ denotes the relation embedding of the pair:

$$r(o_1, o_2) = [o_1; o_2] \quad (1)$$

Let r and r' denote two examples, and let m denote the dimension of the relation embeddings. The Euclidean metric $d(r, r')$ is computed as follows:

$$d(r, r') = \sqrt{\sum_{i=1}^{m} (r_i - r'_i)^2} \quad (2)$$

3.3 Evaluation

Model and Hyperparameters For our BERT model, we use the basic bert-base-uncased model[2], which has 12 layers with a hidden dimension of 768. For 1-nn, we adopt the scikit-learn (Pedregosa et al., 2011) KNeighborsClassifier implementation.

Training and Test Data for Probing Classifier Following similar sizes of other probing datasets (Vulić et al., 2020; Tenney et al., 2019b), we set aside a test set consisting of 1,998 positive (IS-A) examples (generated from 666 3-tuples) and 1,998 negative (not IS-A) examples. We split the remaining examples into 3 equal training sets, each consisting of 999 positive examples (generated from 333 3-tuples) and 999 negative examples. We report the average score over the 3 runs.

For the transitive examples $[(A, B), (B, C), (A, C)]$ in the test set, the average numbers of hops for (A, B) and (B, C) are 1.5 and 2.1 respectively. This difference is due to the fact that the senses with at least an example sentence are not evenly distributed along a path for nouns in WordNet. On average, only 46% of senses on a sampled path have example sentences, out of which 72% of the senses in the bottom half (i.e., the half closer to the leaves) of the path are associated with example sentences, whereas only 17% of the top half have example sentences. Therefore, when a sense C is sampled from the top half of the path, it is likely to be further away from sense B.

Evaluation Metric We adopt accuracy as our evaluation metric, which measures the percentage of test examples correctly predicted by the probing classifier. All accuracy scores are computed using the scikit-learn package.

[2] https://huggingface.co/transformers/pretrained_models.html

Pairs	IS-A		IS-A and not IS-A	
	# examples	Acc.	# examples	Acc.
All	1998	65.2 ± 2.8	3996	72.6 ± 0.9
1-hop	702	65.8 ± 2.4	1404	72.3 ± 0.9
2-hop	560	64.4 ± 3.5	1120	73.2 ± 1.8
3-hop	377	68.1 ± 3.7	754	72.4 ± 0.8
4-hop	212	65.6 ± 2.4	424	74.1 ± 1.1
5-hop	67	60.7 ± 4.6	134	71.6 ± 1.2
6-hop	40	58.3 ± 8.2	80	70.0 ± 4.7

Table 1: Accuracy scores on the test set, in the form of mean ± standard deviation. Columns 2–3: Accuracy (%) scores for n-hop IS-A pairs. Columns 4–5: Accuracy (%) scores for n-hop IS-A pairs and the same number of not IS-A pairs. Since longer paths are fewer and senses with example sentences are fewer when they are more distant from the leaf, the number of sampled pairs becomes fewer as the number of hops increases. Hops more than 6 are not shown as the number of n-hop examples for any $n > 6$ is fewer than 20.

$p(AB)$	$p(BC)$	$p(AC)$	$p(AC\|AB, BC)$
63.1 (1.9)	66.3 (3.9)	66.2 (2.9)	82.4 (3.1)

Table 2: Accuracy (%) scores for the 666 transitive 3-tuples in the test set. The standard deviations across three runs are shown in parentheses.

4 Experimental Results

4.1 Results Grouped by Number of Hops

The accuracy scores for the test set are shown in Table 1. The overall accuracy score for all pairs of both IS-A and not IS-A classes is 72.6%, suggesting that BERT correctly predicts IS-A relations to some extent. We also provide a breakdown of the accuracy scores according to different number of IS-A hops. The scores indicate that BERT predicts IS-A relations with higher accuracy for smaller number of hops (1–4) than for larger number of hops (5–6), although the prediction accuracy does not drop by a large amount when the number of hops increases, and the accuracy does not vary too much within 1–4 hops.

4.2 Prediction Ability for Transitivity

We quantify BERT's prediction ability for transitivity by measuring how often BERT makes logically consistent predictions for IS-A relations. Specifically, suppose word senses (A, B, C) form the following transitive IS-A relations: *A is-a B is-a* C. We measure how often BERT correctly predicts the IS-A relation (A, C) given that it correctly predicts (A, B) and (B, C). Table 2 shows the accuracy scores for the 666 transitive 3-tuples. In the table, $p(AB)$ denotes the percentage of correctly predicted (A, B) in the 666 (A, B) pairs. Similar definitions apply to $p(BC)$ and $p(AC)$. $p(AC|AB, BC)$ denotes the percentage of correctly predicted (A, C), given that (A, B) and (B, C) are correctly predicted. The conditional probability in Table 2 indicates that when BERT predicts that A *is-a* B and B *is-a* C, it correctly predicts that A *is-a* C 82.4% of the time. That A *is-a* C is not always predicted correctly (given that BERT correctly predicts A *is-a* B and B *is-a* C) suggests that BERT lacks the ability to make logically consistent predictions.

5 Conclusion

In this paper, we have investigated how much BERT agrees with the transitivity constraint of the IS-A relation, via a minimalist probing setting. Our findings indicate that although BERT can predict IS-A relations to some extent, it does not always make logically consistent predictions. Allowing BERT and more generally neural network models to enforce the transitivity constraint of the IS-A relation would be a worthy future research goal. Besides the IS-A relation, there are other transitivity relations like after, before, larger than, smaller than, etc. It would also be interesting to investigate to what extent BERT also enforces or fails to enforce these other transitivity relations in future work.

Acknowledgments

The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). We thank the anonymous reviewers for their helpful comments, and Wei Lu for his constructive suggestions and feedback on this work.

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

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*, pages 4171–4186.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. In *Transactions of the Association for Computational Linguistics*, pages 34–48.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*.

Janosch Haber and Massimo Poesio. 2020. Word sense distance in human similarity judgements and contextualised word embeddings. In *Proceedings of the Probability and Meaning Conference*, pages 128–145.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Nora Kassner and Hinrich Schutze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: a review of machine learning interpretability methods. In *Entropy*.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. In *Computational Linguistics*, pages 387–443.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. In *Proceedings of the National Academy of Sciences of the United States of America*, pages 30046–30054.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6792–6812.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, BERT? assessing BERT as a distributional semantics model. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 279–290.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, pages 2825–2830.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473.

Yada Pruksachatkun, Phil Yeres, Haokun Liu, Jason Phang, Phu Mon Htut, Alex Wang, Ian Tenney, and Samuel R. Bowman. 2020. jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 109–117.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7222–7240.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th Conference on Natural Language Processing*, pages 161–170.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176.