It is a Bird Therefore it is a Robin: On BERT's Internal Consistency Between Hypernym Knowledge and Logical Words

Nicolas Guerin
Ecole Normale Superieure, France
nicolas.guerin@polytechnique.edu

Emmanuel Chemla
Ecole Normale Superieure, France
emmanuel.chemla@ens.psl.eu

Abstract

The lexical knowledge of NLP systems should be tested (i) for their internal consistency (avoiding groundedness issues) and (ii) both for content words and logical words. In this paper we propose a new method to test the understanding of the hypernymy relationship by measuring its antisymmetry according to the models. Previous studies often rely only on the direct question (e.g., *A robin is a ...*), where we argue a correct answer could only rely on collocational cues, rather than hierarchical cues. We show how to control for this, and how it is important. We develop a method to ask similar questions about logical words that encode an entailment-like relation (e.g., *because* or *therefore*). Our results show important weaknesses of BERT-like models on these semantic tasks.

1 Introduction

The main training task of transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019) is to predict which word may occur in a given position in a sentence. As a first pass, syntax understanding is an important prerequisite to complete this task through which systems learn the distribution of words within sentences, satisfying the constraints imposed by the linguistic environments these words are in. Accordingly, these models have shown strong syntactic capabilities (Goldberg, 2019; Wu et al., 2020; Warstadt et al., 2019; Jumelet and Hupkes, 2018).

What do they learn about semantics? Hypernymy offers a strong opportunity to study this question as it is very close to entailment, the cornerstone relation in semantics. Also, it can be studied solely through the Masked Language Modelling task, and without fine-tuning. For instance, in the prompt *A robin is a [MASK]*, BERT assigns a high probability to *bird* in the MASK position (Petroni et al., 2019; Jiang et al., 2020). These models have thus captured semantic information about the *relations* between content words, here a relation between *robin* and *bird*. In this work, we begin by following up on the nuanced findings in this area (Hanna and Mareček, 2021; Ravichander et al., 2020), using and refining methods to assess the understanding of hypernymy, pair by pair.

Then we use these initial results and measurements to study the semantics of logical words, and more specifically connectives, such as *thus* or *because*. The idea is to evaluate the internal *coherence* of the system. Specifically, we ask whether NLP models coherently assign a high probability to *thus* in the place of the mask in *This is a robin, [MASK] this is a bird*, exactly in these cases where the pair *robin-bird* is independently (and ungroundedly) registered as a hyponym-hypernym pair.

We thus raise and answer these research questions: Do BERT-like models understand the asymmetric taxonomic relationship of hypernymy (or only a symmetric co-occurrence relation between hypo-hypernyms)? Do they use entailment-like connectives appropriately? Do they show internal consistency: using entailment connectives to connect cases where *they* detect hypernymy (i.e. indepedently of whether hypernymy actually holds)? Hence, our contributions are as follows:

- We test the non-symmetric aspect of hypernymy. To our knowledge, this is absent from other studies, which only test hypernymy through one-sided prompts.
- We extend the methodology to test the semantics of logical connectives like *because* and *therefore*.
- We analyze logical connectives in a non-grounded manner: we test the semantic knowledge of entailment connectives, using entailment facts (hypernyms) that are independently proved to be known by the system.

Findings of the Association for Computational Linguistics: ACL 2023, pages 8807–8817
July 9-14, 2023

- We show that BERT-like models have important weaknesses on all previous tasks. The most surprising one being a reversed semantics for *because*.

2 Semantics As Internal Consistency

One classical approach to semantics is that knowing the meaning of a sentence is knowing in which situations this sentence is true, that is, being able to map (sentence, situation) pairs onto truth-values (Davidson, 1967; Lewis, 1970). Text-only-trained machines surely cannot do so, simply because they only take sentences as inputs, not situations.

However, semantics may also be seen as the graph of all entailment relations between sentences. These entailment relations can follow from taxonomic relations between content words: the fact that all *robins* are *birds* will create entailment relations between sentences (e.g., *John saw a robin* entails *John saw a bird*). Being able to identify these is showing a strong command of the meaning of the words *robin* and *bird*, independently of how these words are grounded in the actual world.

Entailment relations between sentences can also follow from the meaning of the logical words they contain. In a "proof-theoretic" approach, one may even say that this is all there is to the semantics of logical words, which are not grounded: the power to create a consistent net of entailment relations.

Our work is part of this recent vision of the notion of meaning for non-grounded LMs (Piantadosi and Hill, 2022).

3 Related Work

NLP models have been tested for their syntactic abilities (Rogers et al., 2020; Lin et al., 2019; Wu et al., 2020; Goldberg, 2019; Warstadt et al., 2019; Jumelet and Hupkes, 2018; Marvin and Linzen, 2018) for which they obtain strong results, but to a lesser extent for their semantic abilities (Rogers et al., 2020; Balasubramanian et al., 2020; Wallace et al., 2019; Ettinger, 2019) for which they show more fragile performances.

Models such as BERT encode world knowledge (Feldman et al., 2019; Jiang et al., 2020). The first part of our work is a direct follow-up of prompt studies (Liu et al., 2021) targeting knowledge of hypernymy which has been shown to be high but fragile and inconsistent (Petroni et al., 2019; Hanna and Mareček, 2021; Ravichander et al., 2020; Bouraoui et al., 2019). We leverage this knowledge to extend the investigation to logical words.

4 Experiment 1: Content Words

4.1 Metrics

Considering a hyponym-hypernym pair such as (*robin*, *bird*), what probability does BERT assign to the hypernym word *bird* in a MASK position:

$$\mathbb{P}[\mathit{MASK} = \mathit{bird} \mid \mathit{A\ robin\ is\ a\ MASK}] \quad (1)$$

For more than 30% of the pairs, the target hypernym is the top-1 word predicted, and in 80% of the pairs, it is in the top-100 (Petroni et al., 2019). This indicates that BERT recognizes that *robin* and *bird* are likely to co-occur in a sentence. We ask whether the system recognizes that the hyponym-hypernym relation is not symmetric, a critical fact that makes hypernymy a variant of entailment (and not of relevance). We do so by correcting the above probability with the probability of that same hypernym, albeit in the reverse configuration. Thus, we consider the log-ratio of (1) and (2):

$$\mathbb{P}[\mathit{MASK} = \mathit{bird} \mid \mathit{A\ MASK\ is\ a\ robin}] \quad (2)$$

Furthermore, like (Jiang et al., 2020; Hanna and Mareček, 2021), we explore a set of prompts and not just one. For each pair of hyponym-hypernym (h, c) (h the **h**ead and c the **c**lass to which h belongs) we start from a template $\mathit{DET}_1\ h\ \mathit{REL}\ \mathit{DET}_2\ c$, with DET_i determiners (e.g. *the*, *a*, *an*, ϵ) and *REL* an instantiation of the hypernymy relation (e.g. *is*, *is a subclass of*, *is a kind of*, *is a sort of*, *is a type of*). We use the model to compute a score for a set of determiners and relations and then we select the prompt with the highest one (more details in Appendix B, with explanations as to how this optimizes the form of the prompt without *a priori* biasing the final log-ratio scores).

Once the prompt is selected, we compute the following **hypernymy score** σ:

$$\sigma(h, c) := \log \frac{\mathbb{P}[\mathit{MASK} = c \mid \mathit{DET}_1^n\ h\ \mathit{REL}\ \mathit{DET}_2^n\ \mathit{MASK}]}{\mathbb{P}[\mathit{MASK} = c \mid \mathit{DET}_1^d\ \mathit{MASK}\ \mathit{REL}\ \mathit{DET}_2^d\ h]} \quad (3)$$

which should be positive for well-understood pairs. Note that the subscript n and d stands for numerator and denominator respectively as the two are optimized separately. Other formulae are just as natural, such as the σ' presented in Appendix A.

WordNet	T-Rex	ConceptNet	BLESS
0.38 (3.85)	1.14 (3.45)	0.85 (2.76)	0.68 (3.42)

Table 1: Mean (and standard deviation) of the σ scores for content words for BERT-base.

4.2 Multi-token Prediction

Some hyponym-hypernym pairs are made of multi-token expressions. For example, *great ape* is tokenized as two tokens. To overcome this difficulty we use the technique presented in (Feldman et al., 2019) consisting in computing the probability of each token independently and iteratively unmasking the token with the highest probability.

4.3 Knowledge Sources

To build hyponym-hypernym pairs we used the following four different knowledge sources: **WordNet** (Miller, 1995) from which we obtained $110,663$ pairs by selecting synsets connected by the *hyponym* relation, **T-Rex** (Elsahar et al., 2018) using the *subclass of* relation we extracted $2,489$ pairs, **ConceptNet** (Speer and Havasi, 2012) with its *IsA* relation lead to $49,572$ pairs and **BLESS** (Baroni and Lenci, 2011) which contains $1,200$ triplets connected with the *hyper* relation. Note that some pairs are made of rare, if not obscure, words, the **BLESS** corpus aims at minimizing this issue, as it has been filtered through crowd sourcing.

4.4 Results

We conducted all experiments on BERT (Devlin et al., 2019), ELECTRA (Clark et al., 2020), DistilBERT (Sanh et al., 2020) and ALBERT (Lan et al., 2020). The results for BERT-base are given in Table 1 (see Appendix C for the other models). The mean of the scores is always positive ($p < 0.001$). This shows that these models encode the hypernymy relation better than chance. Yet, an average of 45% pairs are encoded in the wrong direction (see Fig. 1 for BERT-base). From a grounded approach of semantics, these are errors. In Experiment 2, we take them as an opportunity to look for traces of strong semantics command, as an internal consistency constraint.

5 Experiment 2: Logical Words

The previous experiment establishes how models capture semantic relations between content nouns. We can use these results to investigate how the same models understand logical words. Concretely, one would expect a high probability for words like *thus*, *so*, *therefore* in the following sentence, and a low probability for words like *because*, *since*, *for* as they encode this entailment the other way around:

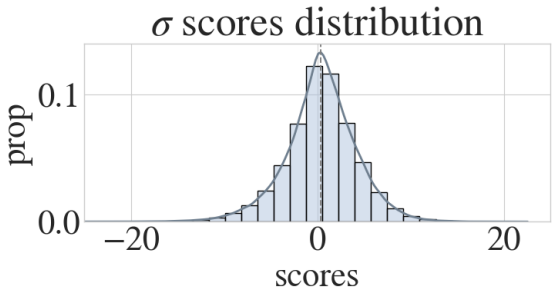

Figure 1: Histogram of the σ score distribution for BERT-base on the WordNet dataset, showing a slightly positive mean (dash-line), as well as a large variance.

It's a robin MASK it's a bird. (4)

Results on hypernym-hyponym pairs show great variability hence, the high probability for *thus*-logical words in the sentence above is expected only if that particular pair, (*robin*, *bird*), is assigned a high hypernymy score by the model. For pairs that receive a very negative score, the expectation is in fact that the results would be reversed. This approach thus allows us to test the semantic *consistency* of the system. Consistency could be perfect for logical words, even if there are grounding errors with content words and world knowledge.

We tested 7 logical words of the *thus* class (*thus*, *therefore*, *consequently*, *then*, *accordingly*, *so*, *hence*), and 5 logical words of the *because* class (*because*, *since*, *for*, *seeing*, *considering*).

5.1 Metrics

We define a score for a logical word w and a hyponym-hypernym pair (h, c) as in (5). This score measures the probability of finding, say, *thus*, in a sentence like (4) above, corrected for the probability of finding it in the reverse sentence.

$$s(w;h,c) := \log \frac{\mathbb{P}[MASK = w \mid PRE_1\ DET_1\ h\ MASK\ PRE_2\ DET_2\ c]}{\mathbb{P}[MASK = w \mid PRE_2\ DET_2\ c\ MASK\ PRE_1\ DET_1\ h]} \quad (5)$$

As before, we explore multiple prompts from a set of determiners *DET* and prefixes *PRE* (see details in Appendix B). A global score $s(l)$ is obtained for a logical word w by averaging $s(w;h,c)$ over the set of all content word pairs (h, c).

As seen in §4.4, the hyponym-hypernym pairs are not all equal regarding to our hypernymy scores. We thus introduce $s^+(w)$ (resp. $s^-(w)$): the average of $s(w;h,c)$ on the top 5%[1] (resp. bottom 5%) pairs according to σ (or σ'). Hence, for a coherent model having understood those logical words we expect $s^+ \geq 0 \geq s^-$ for *thus*-words, and the reverse inequalities $s^+ \leq 0 \leq s^-$ for *because*-words. See Fig. 2 for a graphical representation of the expected results for a consistent model.

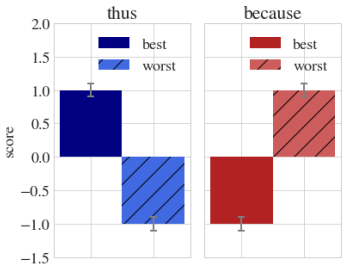

Figure 2: Visual representation of expected results for a consistent model. *best* stands for s^+ and *worst* for s^- We have $s^+ \geq 0 \geq s^-$ for *thus*-words, and the reverse inequalities $s^+ \leq 0 \leq s^-$ for *because*-words.

5.2 Results

Table 2 presents the global scores s for BERT-base (full results are in Appendix C). The *thus*-words almost always obtain a positive global score. The *because*-words sometimes show a negative score (as they should), but most of the times they obtain a positive score just like *thus*-words.

Figure 3 presents the s^+ and s^- scores obtained by BERT-base for the WordNet database relative to the σ score. The *thus*-words obtain a positive score on the best pairs, and a smaller score (albeit not necessarily negative) on the worst pairs. This is the expected result for a model that has correctly understood these logical words. However, *because*-words display a somewhat similar behavior: a positive score over the best pairs, and a lower score over the worst pairs. All models show a qualitatively similar behavior, although ELECTRA seems to behave more consistently (see Appendix C). Overall, these results suggest that *thus*-words and *because*-words alike are understood as being of the *thus*-type.

[1]Empirically we explored several thresholds in percentiles or absolute sigma scores and obtained qualitatively similar results. The 5% threshold was chosen as inclusive enough to have enough pairs to make statistics, and strict enough to make sure the elements in there unambiguously passed the test from Experiment 1.

	WordNet	T-Rex	ConceptNet	BLESS
therefore	0.13	0.12	0.07	0.30
consequently	0.11	0.10	0.04	0.11
then	0.03	0.00	0.00	−0.05
accordingly	0.11	0.07	0.05	0.11
so	0.06	0.04	0.02	−0.03
hence	0.09	0.00	0.05	0.30
thus	0.12	0.00	0.04	0.09
because	0.22	0.26	0.20	0.32
since	0.03	0.11	0.02	−0.23
for	0.01	0.00	0.00	−0.06
seeing	−0.09	−0.13	−0.17	−0.57
considering	−0.03	0.12	−0.02	−0.33

Table 2: Score s for the logical words we tested and for BERT-base. Red numbers represent unexpected results: assuming that content word pairs are well-understood, then a good result would be one with positive scores for the *thus*-words and negative scores for the *because*-words. Here scores for *thus*-words are mainly positive, but they are also positive for the *because*-words.

5.3 Discussion

The similar behavior between *thus* and *because* is puzzling. A first possibility that could explain this would be a significant difference in frequency between *thus* words and *because* words in the training corpus. Indeed a signal that would be too weak for *because* could lead to a poor assimilation of its semantics. Unfortunately we did not check frequencies in the training corpus but according to the python package wordfreq[2], *because* is for example one hundred times more frequent than *therefore* or *thus*, ruling out this explanation. Another possibility is that *because* is not used as the converse of *thus*, even by humans. Underlyingly, the result shows that the sentence *This is a robin, because it is a bird* may be more natural than the reverse *This is a bird, because it is a robin*. One may argue that the latter is somewhat tautological and, as such, not natural, while the former may find its use cases (e.g., when discriminating between a *robin* and an *orangutan*). One may wonder why the converse does not apply to *thus*-words however. To clear this issue one could look at the occurrences of *thus* and *because* in the relevant training corpora. Regardless, a conclusion we can already draw is that the simplest entailment-like semantics for *because* is very far from what is encoded in these models.

6 Conclusion

We propose an approach to the semantic study of BERT-type networks. First we evaluate the models on the non-symmetry of an entailment-like relation, namely hypernymy. The tested models show an

[2]https://pypi.org/project/wordfreq/

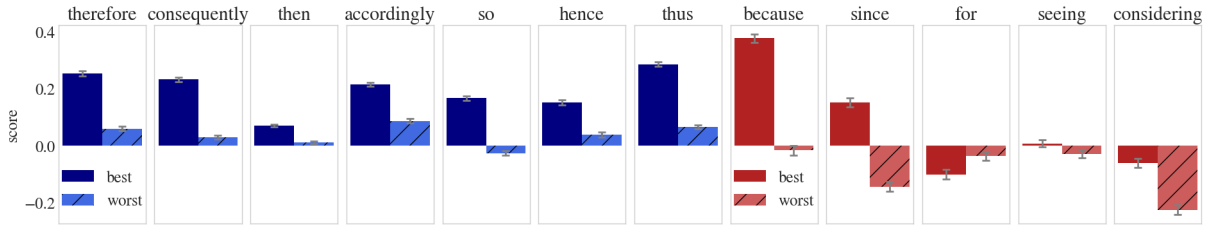

Figure 3: Bar chart of s^+ and s^- scores for BERT-base on the dataset WordNet. The *thus*-like words are in blue, and the *because*-like words are in red. The *thus*-words are relatively well-behaved: the best pairs induce a positive score, while the worst pairs induce a lower score (although not necessarily a negative one, as would be expected). The *because*-words are not as clear: the best pairs also most often yield a positive score, while here we would expect a negative score. Moreover the worst pairs often induce a lower (and always negative) score. Overall, *thus*-words are relatively well understood, while *because*-words are mistaken as *thus*-words.

average positive understanding of this relation. But this is accompanied with a large variance, showing that the relation is very often captured backward.

Thanks to these results we moved to testing logical words of type *thus* and *because*, which impersonate the entailment relation at the core of all enterprises in semantics. Its non-symmetry is one of its fundamental property. The models capture on average the non-symmetry of the words of type *thus* appropriately and they also show good consistency results, that is, a stronger signal for pairs that are themselves well-captured. However, the models obtain similar scores for the words of type *because* and, applying the same standards, they thus capture them backwards. Moreover all these results are to be qualified by their great variability across models and knowledge sources.

These properties albeit simple are yet at the core of what human semantics is however they are not reliably captured. This failure on these basic logical tests then raises questions regarding their otherwise impressive success. They also provide a method to reconstruct their actual semantics, if it is not human-like, and offers challenging tasks for these models.

Limitations

Our evaluations rely on the Masked Language Modelling task as it was a convenient task to conduct our experiments and following up on similar related works. To apply it to models trained differently, e.g., models of the GPT class, one needs to develop comparable appropriateness measures, which is a general desiderata of the field.

We evaluated the models through prompts. We used a fixed set of prompts, and others could produce better results for each of the tasks at stake. Even if one could find *some* working prompts, this may not be ambitious enough, however. It would show that the relevant information is present in the system, say about hyponym-hypernym pairs. But the tests we are proposing rely on the idea that models should work well consistently, across tasks, and across prompts.

With our zero-shot prompting method, we tested the pre-trained models. One could imagine ways to fine-tune these models to our requirements. Our goal here was to first make visible that the groundless training might have been sufficient to encode a consistent semantics, and yet that it did not.

In future work, we also hope to develop quantitative measures of success and consistency (starting from the statistical models in Appendix D), consistency measures which compare parallel performance over more tasks at the same time.

Ethics Statement

We will evaluate and compensate for the carbon footprint of our computations.

Acknowledgements

The research leading to these results was supported by ANR-17-EURE-0017, and was granted access to the HPC resources of GENCI-IDRIS under the allocation AD011013783.

References

Sriram Balasubramanian, Naman Jain, Gaurav Jindal, Abhijeet Awasthi, and Sunita Sarawagi. 2020. What's in a name? are bert named entity representations just as good for any other name?

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2019. Inducing relational knowledge from bert. (arXiv:1911.12753). ArXiv:1911.12753 [cs].

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. ArXiv:2003.10555 [cs] type: article.

Donald Davidson. 1967. Truth and meaning. *Synthese*, 17(1):304–323.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Allyson Ettinger. 2019. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models.

Joshua Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Michael Hanna and David Mareček. 2021. Analyzing bert's knowledge of hypernymy via prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, page 275–282, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-factr: Multilingual factual knowledge retrieval from pretrained language models.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of lstms to understand negative polarity items. pages 222–231.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*.

David Lewis. 1970. General semantics. 22(1–2):18–67.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, and Paul G Allen. 2019. Roberta: A robustly optimized bert pretraining approach.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? pages 2463–2473.

Steven T. Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. (arXiv:2208.02957). ArXiv:2208.02957 [cs].

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in bert. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, page 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretič, and Samuel R Bowman. 2019. Investigating bert's knowledge of language: Five analysis methods with npis.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting bert.

A Another hypernymy score σ'

We introduce another **hypernymy score** σ' that corrects (1) differently:

$$\sigma'(h,c) := \log \frac{\mathbb{P}[\mathit{MASK} = c \mid \text{DET}_1^n\ h\ \text{REL}\ \text{DET}_2^n\ \mathit{MASK}]}{\mathbb{P}[\mathit{MASK} = h \mid \text{DET}_1^d\ c\ \text{REL}\ \text{DET}_2^d\ \mathit{MASK}]} - \log \frac{\mathbb{P}[\mathit{MASK}_2 = c \mid \text{DET}_1^d\ \mathit{MASK}_1\ \text{REL}\ \text{DET}_2^d\ \mathit{MASK}_2]}{\mathbb{P}[\mathit{MASK}_2 = h \mid \text{DET}_1^n\ \mathit{MASK}_1\ \text{REL}\ \text{DET}_2^n\ \mathit{MASK}_2]} \quad (6)$$

There, the first log-ratio involves the probability of the hyponym in the reverse sentence. We also control for the mere probabilities of hypernyms and hyponyms without the influence of the other (by masking the other), which is the role of the second log-ratio. We report results for both of these scores.

B Prompts

To compute the prompts we chose a zero-shot approach to model probing. Starting from a template of the form:

$$\mathit{DET}_1\ \mathit{NOUN}_1\ \mathit{REL}\ \mathit{DET}_2\ \mathit{NOUN}_2$$

DET_i are determiners that we can chose from the set $\{$*the*, *a*, *an*, $\epsilon\}$, *REL* is an instantiation of the hypernymy relation that can be $\{$*is*, *is a subclass of*, *is a kind of*, *is a sort of*, *is a type of*$\}$ and NOUN_i are placeholders for the hyponyms-hypernyms and for the *MASK* token during inference.

Then for each pair (h, c) we computed an optimal REL^* that maximizes:

$$\max_{\text{DET}_1^n, \text{DET}_2^n} \Big(\mathbb{P}(\text{MASK}_2 = c \mid \mathit{DET}_1^n\ \mathit{MASK}_1\ \mathit{REL}\ \mathit{DET}_2^n\ \mathit{MASK}_2) \times \mathbb{P}(\text{MASK}_1 = h \mid \mathit{DET}_1^n\ \mathit{MASK}_1\ \mathit{REL}\ \mathit{DET}_2^n\ \mathit{MASK}_2) \Big) \times \max_{\text{DET}_1^d, \text{DET}_2^d} \Big(\mathbb{P}(\text{MASK}_2 = c \mid \mathit{DET}_1^d\ \mathit{MASK}_2\ \mathit{REL}\ \mathit{DET}_2^d\ \mathit{MASK}_1) \times \mathbb{P}(\text{MASK}_1 = h \mid \mathit{DET}_1^d\ \mathit{MASK}_2\ \mathit{REL}\ \mathit{DET}_2^d\ \mathit{MASK}_1) \Big) \quad (7)$$

By selecting the determiners that realize the max in the previous equation which are DET_i^{n*} for the numerator of the hypernymy score (3) and DET_i^{d*} for the denominator, one obtain two prompts:

$$\mathit{DET}_1^{n*}\ \mathit{NOUN}_1\ \mathit{REL}^*\ \mathit{DET}_2^{n*}\ \mathit{NOUN}_2$$

and

$$\mathit{DET}_1^{d*}\ \mathit{NOUN}_1\ \mathit{REL}^*\ \mathit{DET}_2^{d*}\ \mathit{NOUN}_2$$

Note that we constrain the relation instantiation *REL* to be the same in both prompts to gain computation time.

The idea behind (7) is to investigate for which prompt the model prefers to favor the probability of the hypernym or the hyponym in any position in the sentence but without the influence of the other word of the pair (hence two *MASK*s, where the placeholders NOUN_i were). This encourages the selection of an appropriate prompt independently of the overall truth-value of the sentence with the two words, which is precisely what we want to study afterwards.

For logical words the idea is exactly the same. We explore a set of prompts with DET_i determiners from $\{$*the*, *a*, *an*, $\epsilon\}$ and PRE_i prefixes from $\{$*this is*, *it is*$\}$ chosen to maximize for a given logical word w and a given pair (h, c):

$$\mathbb{P}[\mathit{MASK} = w \mid \mathit{PRE}_1\ \mathit{DET}_1\ h\ \mathit{MASK}\ \mathit{PRE}_2\ \mathit{DET}_2\ c] \times \mathbb{P}[\mathit{MASK} = w \mid \mathit{PRE}_2\ \mathit{DET}_2\ c\ \mathit{MASK}\ \mathit{PRE}_1\ \mathit{DET}_1\ h] \quad (8)$$

Here however we constrain the determiners DET_i to be the same in the denominator and the numerator to gain computation time.

C Results

In this section we give the full results across the various datasets and models. We will not conduct full analyses of all results, although results are very heterogenous (even for the same model across different knowledge sources). Table 3 gives the full results for hypernymy scores σ and σ'. Tables 4, 5, 6, 7 and 8 show the logical scores results. The undesirable scores are shown in red.

Impressionistically, ELECTRA seems to have the best consistency results (Fig. 4). At the other end of the spectrum, although it looks like BERT has sometimes well captured *thus*-type logical words, there are also counter-examples to this, on BLESS for instance (Fig. 5).

D Statistics

To obtain a more quantitative measure of the phenomenon, we fit a Linear Mixed Effect Model to evaluate the impact on the logical score of the type of pair, better or worse (ie. top 5% vs bottom 5%) as well as the interaction of this variable with the

		WordNet	T-Rex	ConceptNet	BLESS
BERT-base	σ	0.38 (3.85)	1.14 (3.45)	0.85 (2.76)	0.68 (3.42)
	σ'	1.13 (4.57)	0.98 (4.67)	0.79 (3.26)	0.11 (3.26)
BERT-large	σ	0.79 (3.89)	1.58 (4.00)	1.07 (3.12)	1.62 (2.99)
	σ'	1.28 (5.10)	1.74 (5.53)	1.14 (3.84)	0.91 (3.30)
DistilBERT	σ	0.63 (2.56)	1.03 (2.37)	0.51 (2.18)	1.31 (2.05)
	σ'	0.54 (2.66)	0.52 (2.79)	0.50 (2.22)	0.53 (1.92)
ALBERT	σ	0.63 (3.02)	0.63 (2.97)	0.45 (2.76)	1.09 (1.96)
	σ'	0.78 (9.59)	2.69 (10.15)	0.56 (7.73)	1.86. (8.82)
ELECTRA-small	σ	0.52 (2.06)	0.43 (2.14)	0.51 (1.94)	1.34 (2.01)
	σ'	0.24 (2.55)	0.27 (2.72)	0.17 (2.22)	0.43 (1.71)

Table 3: Mean (and standard deviation) of the scores for content words for the different models and datasets.

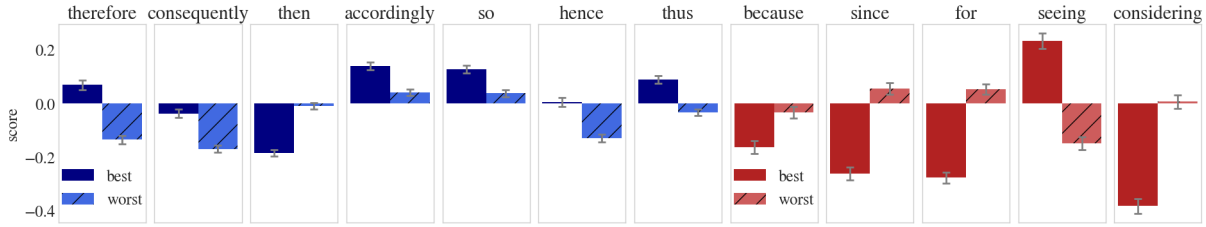

Figure 4: Bar chart of s^+ and s^- scores on the dataset ConceptNet and computed from ELECTRA generator. This is qualitatively the most consistent behavior observed across models and knowledge source.

	WordNet	T-Rex	ConceptNet	BLESS
therefore	0.13	0.12	0.07	0.30
consequently	0.11	0.10	0.04	0.11
then	0.03	0.00	0.00	−0.05
accordingly	0.11	0.07	0.05	0.11
so	0.06	0.04	0.02	−0.03
hence	0.09	0.00	0.05	0.30
thus	0.12	0.00	0.04	0.09
because	0.22	0.26	0.20	0.32
since	0.03	0.11	0.02	−0.23
for	0.01	0.00	0.00	−0.06
seeing	−0.09	−0.13	−0.17	−0.57
considering	−0.03	0.12	−0.02	−0.33

Table 4: Score s for BERT-base.

	WordNet	T-Rex	ConceptNet	BLESS
therefore	0.07	0.16	0.02	0.25
consequently	0.03	0.07	−0.05	0.05
then	−0.07	−0.14	−0.10	−0.22
accordingly	−0.01	0.05	−0.04	0.16
so	0.09	0.09	0.09	0.19
hence	0.09	0.00	0.05	0.30
thus	0.08	0.15	0.01	0.17
because	0.21	0.25	0.16	−0.26
since	0.27	0.24	0.06	−0.38
for	0.11	0.13	0.00	−0.14
seeing	−0.18	−0.15	−0.08	0.22
considering	0.06	−0.11	−0.06	−0.40

Table 5: Score s for BERT-large.

	WordNet	T-Rex	ConceptNet	BLESS
therefore	−0.06	−0.07	0.08	0.31
consequently	−0.02	−0.01	−0.03	−0.04
then	−0.11	−0.08	−0.13	−0.36
accordingly	0.01	0.05	−0.01	0.00
so	0.07	0.04	0.08	0.12
hence	−0.04	−0.09	−0.05	−0.04
thus	0.07	0.03	0.12	0.24
because	−0.11	−0.18	0.09	0.29
since	−0.13	−0.15	0.06	0.06
for	−0.25	−0.31	−0.04	−0.15
seeing	0.10	0.02	0.08	0.12
considering	−0.22	−0.15	0.10	0.15

Table 6: Score s for DistilBERT.

	WordNet	T-Rex	ConceptNet	BLESS
therefore	0.00	0.09	−0.02	−0.17
consequently	−0.05	0.05	−0.06	−0.15
then	0.00	0.11	0.03	−0.10
accordingly	−0.12	−0.06	−0.14	−0.21
so	0.10	0.15	0.09	0.01
hence	0.01	0.11	−0.02	−0.16
thus	0.03	0.11	0.02	−0.05
because	0.10	0.13	0.10	−0.13
since	0.11	0.14	0.11	−0.03
for	0.04	0.07	0.01	−0.19
seeing	0.00	−0.06	−0.03	−0.09
considering	−0.02	−0.06	−0.02	−0.17

Table 7: Score s for ALBERT.

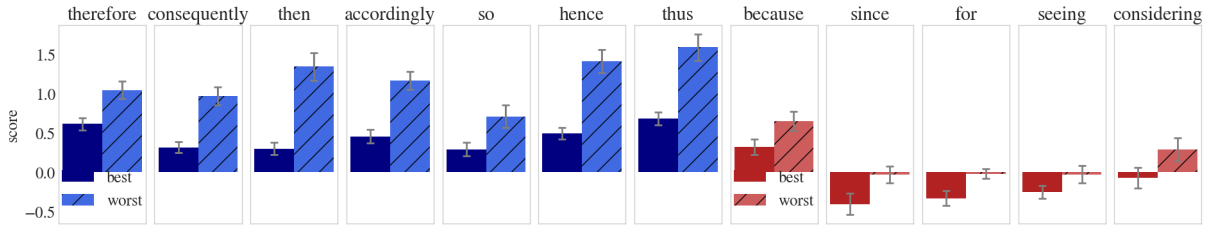

Figure 5: Bar chart of s^+ and s^- scores for BERT-base on the dataset BLESS. Here we have evidence of a more *because*-type understanding of *thus*-type logical words.

	WordNet	T-Rex	ConceptNet	BLESS
therefore	0.13	0.12	0.00	0.23
consequently	0.02	0.00	−0.08	−0.06
then	−0.01	0.04	−0.06	−0.04
accordingly	0.16	0.13	0.11	0.06
so	0.17	0.15	0.11	0.35
hence	0.06	0.05	−0.04	−0.02
thus	0.14	0.12	0.06	0.20
because	−0.02	0.00	−0.09	−0.50
since	0.03	0.04	−0.04	−0.76
for	−0.04	−0.05	−0.07	−0.49
seeing	−0.04	−0.09	0.01	0.01
considering	−0.06	0.01	−0.11	−0.90

Table 8: Score s for ELECTRA-small.

category of the logical word *thus*-like vs. *because*-like. This gives us the following model:

$$Y_{pw} = \beta_0 + S_{0l} + I_{0p} + \beta_p X_p + \beta_l X_w + \beta_{pl} X_p X_w \quad (9)$$

with $X_p = 1$ if the pair p is of type "better" and 0 if it is of type "worse", $X_w = 1$ if the logical word w is *thus*-like and 0 if it is *because*-like, $(\beta_0, \beta_p, \beta_w)$ the fixed effect parameters, β_{pw} the interaction parameter and $S_{0w}, \sim \mathcal{N}(0, \omega_{00}^2)$ and $I_{0p}, \sim \mathcal{N}(0, t_{00}^2)$ the random effect parameters.

We do it for BERT (base and large), excluding BLESS because it has too few pairs. A distinct understanding of *thus*-words vs. *because*-words would be indicated by a positive β_{pw}, the interaction between the category of the pair and the category of the logical words. But this parameter is always either statistically negative or non statistically different from zero. This reveals that *because*-words were not understood differently from *thus*-logical words.

The parameter β_p is 8 times out of 12 significantly positive (3 datasets WordNet, T-Rex and ConceptNet, 2 models BERT-base and BERT-large and 2 scores σ and σ'). In other words, for BERT the top hypernym-hyponym pairs received higher scores than the bottom pairs, suggesting that logical words, *thus* and *because* alike, were understood like *thus*-words.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section: Limitations

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Secation: Abstract and 6. Conclusion

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Section 4 and 5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Because we used well-known BERT model, cited its original paper and only for inference, hence no GPU training.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and 5 (and Appendix)

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.