On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT

Abhilasha Ravichander[♢]*, **Eduard Hovy**[♢], **Kaheer Suleman**[♡]
Adam Trischler[♡♠], **Jackie Chi Kit Cheung**[♠]
[♢]Carnegie Mellon University, Pittsburgh, PA
[♡]Microsoft Research, Montreal, Canada [♠]McGill University, Montreal, Canada
{aravicha, ehovy}@cs.cmu.edu
{adam.trischler, kasulema}@microsoft.com
{jcheung}@cs.mcgill.ca

Abstract

Contextualized word representations have become a driving force in NLP, motivating widespread interest in understanding their capabilities and the mechanisms by which they operate. Particularly intriguing is their ability to identify and encode conceptual abstractions. Past work has probed BERT representations (Devlin et al., 2019) for this competence, finding that BERT can correctly retrieve noun hypernyms in cloze tasks. In this work, we ask the question: *do probing studies shed light on systematic knowledge in BERT representations?* As a case study, we examine hypernymy knowledge encoded in BERT representations. In particular, we demonstrate through a simple consistency probe that the ability to correctly retrieve hypernyms in cloze tasks, as used in prior work, does not correspond to systematic knowledge in BERT. Our main conclusion is cautionary: even if BERT demonstrates high probing accuracy for a particular competence, it does not necessarily follow that BERT 'understands' a concept, and it cannot be expected to systematically generalize across applicable contexts.[1]

1 Introduction

Hierarchical representations of concepts play a central role in reasoning and understanding natural language (Wellman and Gelman, 1992). They have long been studied as a core NLP objective in their own right through tasks requiring the identification of hypernyms (Hearst, 1992; Snow et al., 2005, 2006), and as components for use in downstream

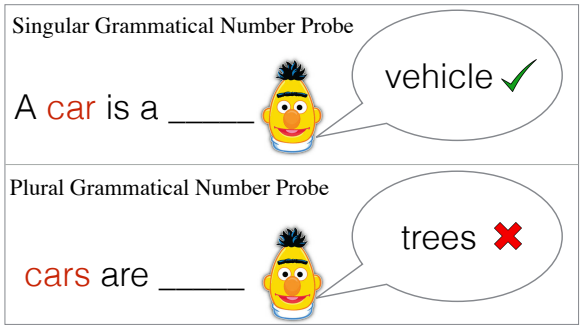


Figure 1: Illustration of BERT's inconsistent predictions on singular and plural hypernymy probes.

NLP tasks, such as recognizing textual entailment (RTE), metaphor detection, text generation and question answering (QA) (Girju et al., 2003; Dagan et al., 2006; Prager et al., 2008; Mirkin et al., 2009; Akhmatova and Dras, 2009; Mohler et al., 2013; Biran and McKeown, 2013; Yahya et al., 2013). Recently, Pretrained Language Models (PLMs), such as BERT (Devlin et al., 2019), have emerged as a popular and successful approach to a variety of NLP tasks. Thus, there has been community interest in evaluating their representations for the 'knowledge' they contain, including information about concept abstraction (Ettinger, 2020; Talmor et al., 2019; Jiang et al., 2020; Petroni et al., 2019).

We distinguish research that investigates knowledge encoded in BERT through two broad perspectives: *instrumentative* and *agentive*. We view the instrumentative perspective as treating PLMs *as a tool* to mine or store knowledge, like hypernym-hyponym and other relations, from text (Petroni et al., 2019; Jiang et al., 2020; Bouraoui et al., 2019; Bosselut et al., 2019; Madaan et al., 2020). The primary purpose of these investigations is to identify effective techniques to extract information from PLMs for use in downstream pipelines. In contrast, a growing body of work adopts an agentive perspective (Ettinger et al., 2018; Talmor et al., 2019),

* Part of this work was done during an internship at Microsoft Research.

[1]Diagnostic framework available at https://github.com/AbhilashaRavichander/probe-generalization.

*Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 88–102
Barcelona, Spain (Online), December 12–13, 2020

treating PLMs as Artificial Intelligence (AI) agents and analyzing their linguistic competencies and world knowledge, sometimes through tasks such as natural language inference (Williams et al., 2018; Wang et al., 2018) or story completion (Zellers et al., 2018, 2019; Mostafazadeh et al., 2016).

In this work, we examine the agentive perspective, focusing specifically on the validity of conclusions drawn from probing studies. A popular approach to probing knowledge in pre-trained language models is the *zero-shot masked-LM* probing task. For example, given the statement 'A robin is a [MASK]', a PLM that produces the correct completion 'bird' is considered successful.[2] Past work has studied this competency in BERT (Ettinger, 2020), offering BERT's ability to correctly retrieve noun hypernyms in cloze tasks as evidence that it successfully encodes hypernymy information.

But to what extent does this knowledge of hypernymy generalize? Among many systematic generalization abilities desirable in PLMs, we select the following two. (1) Syntagmatic generalization: A model that has knowledge of a fact will be able to correctly answer queries about it and apply it across different contexts; (2) Paradigmatic generalization: A model with a particular competency will be able to generalize to novel cues and items. We implement these generalization requirements through a set of diagnostic probing tasks, in which a model must demonstrate consistency in applying its knowledge across different selective contexts, and by generalizing in trained probe settings to novel, unseen items belonging to the same semantic category or relation.

In particular, we focus on the setting of Ettinger (2020), which demonstrates that BERT is "very strong at associating nouns with hypernyms." We propose consistency tasks to illuminate the limits and generality of this ability, as illustrated in Figure 1. Our consistency tasks combine related zero-shot probes in such a way that a model that succeeds on one probe, if it is drawing on a systematic, general ability, should also succeed on the paired probe. Our evaluation with a grammatical number consistency task sheds light on the fragility of BERT's ability to associate correct noun hypernyms and demonstrates that pre-trained LMs have considerable room for improvement to reach a human-like level of understanding. [3]

Contributions: We demonstrate success on a hypernymy probing benchmark does not necessarily correspond to a systematic conceptual understanding of the phenomena in BERT, as discovered by probes. We further formulate evaluation protocols for characterizing the generalizability of PLM knowledge, in order to draw more reliable conclusions from probing studies.[4]

2 Experimental Methodology

Saussure (1916) expounds on syntagmatic relations, studying how words acquire relations based on the ways in which they are chained together in language context. The syntagmatic relation is based on groups of terms, in this case the hyponym and hypernym that are communicated together. In this work, we study whether PLM probes generalize syntagmatically, by evaluating the ability of models to produce correct predictions for hyponym-hypernym items across both singular and plural contexts. We also examine the ability of probes to generalize *paradigmatically*, that is, do probing studies uncover paradigms embedded in text (in this case the relations between items and their abstractions)?[5]

2.1 Syntagmatic Generalization

Knowledge in BERT is often studied using zero-shot probes (Ettinger, 2020; Talmor et al., 2019) in a *masked LM format*. In this construction, a PLM is queried by a natural language prompt designed to exercise a particular competence; for example, '*A robin is a [MASK]*' to evaluate knowledge of hypernymy. The word assigned the highest probability at the masked position is considered the PLM's answer.

In this work, we design diagnostics to examine how systematically this "knowledge" generalizes. We consider two kinds of diagnostics—(1) Consistency: We evaluate a PLM's ability to consistently answer queries reflecting the same conceptual understanding. We use a simple number consistency

[2] We refer to such probes henceforth as zero-shot masked LM probes, since they require no training and use BERT's masked-LM component to fill in the answer.

[3] Consistency tasks can be considered complementary to the control tasks proposed by Hewitt and Liang (2019). While control tasks test attribution, consistency tasks test validity.

[4] Our study is based on probes in English.

[5] This distinction is concerned with the axis of generalization of probes. In our syntagmatic generalization probes, we are concerned with different lexico-syntactic contexts where a model can demonstrate its knowledge of hypernymy. In the paradigmatic generalization probes, we are concerned with generalizing to novel hypernym/hyponym pairs.

check for hypernymy. Queries with hyponyms are replaced by their plural forms ; e.g. '*A robin is a [MASK]*' is perturbed as '*robins are [MASK]*'. Agents drawing on a general taxonomic reasoning ability should be able to correctly answer queries in both forms. (2) Contextual: We examine a PLM's ability to recognize the correct abstraction for a hyponym in context; e.g. '*A robin perches in its nest.*' is replaced with '*A [MASK] perches in its nest.*', where the hypernym *bird* is an acceptable substitution. Agents that understand concept instantiations should identify the correct abstraction.

2.1.1 Probes

Consistency Probes: In this paper, we adopt a *zero-shot cloze formulation* where hypernymy knowledge is in the form of triples $< x, y, t_{1...n} >$. Here, x is a hyponym, y is a hypernym and t_i is a cloze-style prompt consisting of a sequence of tokens, two of which are placeholders for the hyponym and hypernym (e.g., "A x is a y"). The final probe replaces x with the surface form of the hyponym, and lets the model predict the missing hypernym y (e.g., 'A robin is a ___').

Contextualized Probes: We further define a *contextual probe* formulation, wherein hypernymy knowledge is in the form of triples $< x, y, t_{1...n} >$. Here, x is a hyponym, y is a hypernym and t_i is a sequence of tokens, one of which is the hyponym y. The final probe asks the model to predict an appropriate hypernym x in the place of y in t_i. (e.g.., 'A ___ perches in its nest.').

2.1.2 Datasets

LM DIAGNOSTIC: We use the NEG-136 diagnostic constructed by Ettinger (2020), selecting the affirmative contexts to test models' use of hypernym information. Test items are drawn from a human study conducted by Fischler et al. (1983), wherein subject words are 18 concrete nouns and hypernyms belong to nine superordinate categories (Battig and Montague, 1969).[6] The final diagnostic set consists of 18 prompts.

LM DIAGNOSTIC EXTENDED: In this work, we additionally expand LM DIAGNOSTIC to construct a larger diagnostic set. For each superordinate category (Battig and Montague, 1969), we extract hyponyms from WordNet (Fellbaum, 1998a) such that they are nouns, not named entities, and only have a single sense in WordNet. This enables us to construct an expanded diagnostic set of 576 prompts. Statistics of both datasets, as well as sample queries, are reported in Table 1.

For each query in both datasets, we construct grammatical number consistency probes. Each query is perturbed to contain both the subject hyponym and target hypernym in plural form. Additionally, we construct contextual probes for each subject hyponym and target hypernym. These manually-crafted probes examine a PLM's abilities to identify correct abstractions for concepts in context. Each query consists of a sentential context collected from Wikipedia that contains the *hyponym* but not the hypernym, so as not to give easy cues to the LM. Each sentential context also satisfies the following additional requirements: (a) permissive of the abstraction (for example, the context *"The New York Public Library was built in the 1890's"* permits the *building* abstraction, but *"The New York Public Library fired John"* does not), (b) selective of the correct hypernym (for example, the context of the target item 'robin' in *"The charity began preservation efforts to save the robin"* is applicable to other categories besides the correct hypernym category *'bird'*—such as the *'insect'* category), and (c) upward entailing of the correct hypernym abstraction (for example, *"The largest salmon caught in the lake was 150cm"* does not entail *"The largest fish caught in the lake was 150cm"*).

2.2 Paradigmatic Generalization

We also examine conceptual generalization of the hypernymy relations: does hypernymy present a systematic pattern in the contextualized embedding space that enables generalization to novel items? To study this, we follow the popular probing methodology of training classifiers to predict hypernym relations from contextualized representations, with no task-specific fine-tuning.

Broadly, the task can be defined as follows. Given a pair of words a_1 and a_2, each grounded in a sentential context, s_1 and s_2, respectively, the goal is to describe whether a_1 and a_2 are in a hypernymy relation. For example, <building> is a hypernym of <skyscraper>, but <vehicle> is not. To examine generalization, we construct probing datasets with two settings: one where hypernyms are seen during training but hyponyms remain entirely unseen (SEEN), and one where both hyponyms and hypernyms in the tests are unseen during training

[6] bird, insect, fish, vehicle, tool, building, tree, flower, vegetable

(UNSEEN). All datasets are constructed to enable three-fold cross-validation.[7] In all cases, each train instance is provided with multiple contexts from Wikipedia but test sets only feature one context per hyponym-hypernym pair.

2.2.1 Probes

We follow the work on diagnostic classifiers (Shi et al., 2016; Adi et al., 2017; Conneau et al., 2018; Hupkes et al., 2018; Liu et al., 2019; Shwartz and Dagan, 2019) and construct minimal embed-interact-predict probes to assess taxonomic knowledge in pretrained representations.

Embed: We embed each word in the hypernymy pair using the embedding model to obtain $\langle w_1, w_2 \rangle$. These representations can either be functions of the word itself (in static embeddings) or functions of the entire sentence (in contextualized embeddings).

Interact: Following Vu and Shwartz (2018), we concatenate the representations w_1, w_2 with their difference $w_2 - w_1$, and their element-wise product $w_1 \odot w_2$ to form representation $\vec{x}$.

Predict: We then apply a softmax classifier over the formed representation-
$\vec{o} = softmax(W \cdot ReLU(Dropout(h(\vec{x}))))$
where h is a 300-dimensional hidden layer, dropout probability = 0.2, $W \in \mathcal{R}^{n \times 300}$, and n=2.

2.2.2 Datasets

We select hyponym-hypernym pairs from LM DIAGNOSTIC EXTENDED. For each dataset, we pair both the hyponym and the hypernym with sentential contexts from Wikipedia.[8] We construct challenging negative examples by choosing hypernyms that belong to the same superordinate category [9] and which are not hypernyms of the word itself. We construct the datasets to meet the following specifications: (1) All datasets are balanced so that simple accuracy can be used as an evaluation metric, (2) Target pairs do not appear across train/test partitions to mitigate lexical memorization (Levy et al., 2015), (3) Negative examples should be similar words, so that simply exploiting distributional similarity does not work, (4) All examples are grounded in phrasal or sentential context.

[7] Statistics of these datasets can be found in the appendix, Table 5 and Table 6.

[8] For both hyponyms and hypernyms, contextualized word representations are extracted using 'context embeddings' (Coenen et al., 2019). The input to BERT is a sequence of tokens from the sentential context and the output consists of a sequence of vectors corresponding to the input tokens. To obtain a representation for a hyponym or hypernym in a sentential context, we construct the average of the output vectors for the tokens in the hyponym or hypernym.

[9] animals, plant, object

3 Syntagmatic Generalization

3.1 Metrics

We consider the following rank-based metrics:

Open vocabulary accuracy: We compute mean precision@k (Open Voc.) where for a given hyponym, the value is 1 if the hypernym is ranked in the top k results and 0 otherwise. We report results with both $k = 1$ and $k = 5$. In the open vocabulary setting, the candidate list is BERT's vocabulary.

Singular accuracy: For a given hyponym, the query is posed in the singular form (e.g., *'A robin is a [MASK]'*), and PLMs are evaluated on their ability to identify the correct hypernym from the nine Fischler categories, where the category assigned the highest probability by the PLM is considered the answer, as in prior work. The value is 1 if the correct hypernym is the top result and 0 if not.

Plural accuracy: For a given hyponym, the query is posed in the plural form (e.g., *'Robins are [MASK]'*), and PLMs are evaluated on their ability to identify the correct hypernym from the nine Fischler categories in plural form, where the category assigned the highest probability by the PLM is considered the answer. The value is 1 if the correct hypernym is the top result and 0 if not.

Contextual accuracy: For a given hyponym, PLMs are evaluated on their ability to identify the correct hypernym in context, evaluated over the nine Fischler categories in singular form.

Paired Singular-Plural accuracy: For a given hyponym item, PLMs are evaluated on their ability to identify the correct hypernym in both singular and plural probes, over a candidate space of the nine Fischler categories. The value is 1 if the correct hypernym is the top answer in both cases.

Paired Aggregate accuracy: For hyponyms with a contextual probe, PLMs are evaluated on their ability to identify the correct hypernym in singular, plural and contextual probes, evaluated over the nine Fischler categories. The value is 1 if the correct answer is the top answer in all three cases.

3.2 Baselines and Models

We compare to the following baselines:

Dataset	Format	# Examples	Example
LM DIAGNOSTIC (Ettinger, 2020)	Zero-shot Cloze	18	A robin is a [MASK]
LM DIAGNOSTIC EXTENDED Singular	Zero-shot Cloze	576	A robin is a [MASK[
LM DIAGNOSTIC EXTENDED Plural	Zero-shot Cloze	576	Robins are [MASK[
LM DIAGNOSTIC EXTENDED Contextual	Zero-shot Cloze	186	Through use of an awl [TOOL] , the surgeon creates tiny fractures in the subchondral bone plate

Table 1: Statistics of zero-shot cloze probing datasets to study syntagmatic generalization.

Model	Open Voc. k=1	Open Voc. k=5	Singular	Plural	Contextual	Paired Singular-Plural	Paired Aggregate
LM DIAGNOSTIC							
Majority	-	-	11.11	11.11	11.11	11.11	11.11
word2vec	0.0	50.0	83.33	100.0	-	83.33	-
GloVe	0.0	27.78	88.89	100.0	-	88.89	-
FastText	0.0	0.0	22.22	16.67	-	0.0	-
BERT-control	0.0	11.11	44.44	55.56	-	38.89	-
BERT	38.89	100.0	**100.0**	77.78	66.67	**77.78**	50.0
LM DIAGNOSTIC EXTENDED							
Majority	-	-	22.92	22.92	31.72	22.92	31.72
word2vec	3.47	18.06	60.59	54.69	-	43.75	-
GloVe	0.35	3.3	58.16	50.17	-	35.24	-
FastText	0.0	0.0	12.15	11.11	-	1.91	-
BERT-control	0.35	2.08	30.56	39.76	-	20.66	-
BERT	23.09	48.96	**67.53**	44.1	73.66	**36.63**	33.33

Table 2: Performance of models on syntagmatic generalization probes. In the open Voc. k=1 and open Voc. k=5, we report mean precision@k, when the candidate list is BERT's vocabulary. We report accuracy(%) for singular, plural and contextual probes, where the candidate list is the nine superordinate categories (Battig and Montague, 1969)-bird, insect, fish, vehicle, tool, building, tree, flower, vegetable- in singular, plural and singular forms respectively. Paired singular-plural accuracy(%) is performance on identifying the correct hypernym in both singular and plural probes. Paired aggregate accuracy(%) is performance on identifying the correct hypernym in singular, plural and contextual probes, if a contextual probe for the hyponym exists.

Majority: Simple majority baseline quantifying the performance of a model that always predicts the majority class in the test set.

Static embedding: For each hyponym, we extract the static embedding with minimum cosine distance to the embedding of the hyponym word, amongst the Fischler categories. We evaluate the following word embeddings. (1) word2vec (Mikolov et al., 2013): Word embeddings are the hidden representations of a feedforward network trained to predict words in a fixed surrounding window to a particular word. We use the 300-dimension English word vectors trained on the Google News corpus. (2) GloVe (Pennington et al., 2014): GloVe embeddings are generated through training models to estimate the log-probability of word-pair co-occurence. We use 300-dimensional GloVe vectors trained on 6B tokens of text. (3) FastText (Bojanowski et al., 2017): FastText vectors extend word2vec with sub-word information. We use 300-dimensional vectors trained on Wikipedia.

BERT-control (Devlin et al., 2019): Following Talmor et al. (2019), we define a simple BERT control which does not include relation information in the probe. Each query consists of the hyponym word followed by the *'[MASK]'* token (e.g., *'robin [MASK]'*) and the probability assigned by the PLM to the candidate list is computed.

BERT (Devlin et al., 2019): Bidirectional Encoder Representations from Transformers (BERT) is based on the transformer architecture (Vaswani et al., 2017) and is trained with both a cloze-style and next-sentence prediction objective.

3.3 Results

Table 2 displays performance scores of BERT on zero-shot probing tasks. We observe that in agreement with prior work, BERT achieves impressive results on the LM DIAGNOSTIC dataset in the open vocabulary setting, providing the right hypernym as the top answer for 38.89% of samples, and within the top 5 answers for 100.0% of samples. However, the LM DIAGNOSTIC consists of only 18

	Open	Singular	Plural	
Prompt	Predictions	Predictions	Prompt	Predictions
LM DIAGNOSTIC				
A robin is a [MASK]	**robin**, bird, pigeon	**bird**, flower, tree	**robins are [MASK]**	**flowers**, birds, trees
A trout is a [MASK]	fish, trout, fishery	fish, bird, tool	trout are [MASK]	fish, trees, birds
A car is a [MASK]	**car**, vehicle, driver	**vehicle**, building, tool	**cars are [MASK]**	**trees**, vehicles, fish
LM DIAGNOSTIC EXTENDED				
An aircraft is a [MASK]	**glider**, helicopter, aircraft	**vehicle**, bird, building	**aircraft are [MASK]**	**fish**, trees, buildings
A bumblebee is an [MASK]	**insect**, animal, airplane	**insect**, bird, flower	**bumblebees are [MASK]**	**birds**, insects, flowers
A bedbug is an [MASK]	**animal**, insect, object	**insect**, tool, vegetable	**bedbugs are [MASK]**	**fish**, flowers, insects

Table 3: Examples of BERT predictions for hypernymy relations with divergences highlighted in red, and samples with inconsistent predictions **in bold**. In the open vocabulary setting, the candidate list is BERT's vocabulary. In the singular probe setting, the candidate list is the nine superordinate categories from (Battig and Montague, 1969). In the plural setting, the candidate list is the nine categories from (Battig and Montague, 1969) in plural form, and the query is converted to the plural form.

such queries, and we observe that this performance drops considerably on the expanded diagnostic dataset LM DIAGNOSTIC EXTENDED (N=576), with the right hypernym being the top answer for only 23.09% of samples and within the top 5 answers for only 48.96% of samples.[10] We further observe that in both diagnostic datasets, BERT performance scores on plural probes are often lower than singular probes. The examples answered correctly in both plural and singular form in the LM DIAGNOSTIC EXTENDED dataset constitute approximately *half* what a standard singular zero-shot probe might lead a practitioner to believe. This is problematic, since if BERT possesses systematic 'knowledge' as discovered by probes, it ought to generalize in robust ways across our diagnostics.

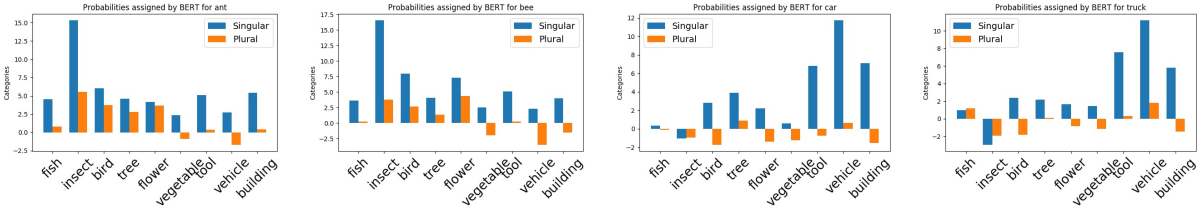
Table 3 features BERT predictions on both diagnostic datasets, with divergences highlighted. We observe that in the open vocabulary setting, BERT predicts correct abstractions not included within the LM DIAGNOSTIC categories. To further estimate the kinds of errors that occur in BERT predictions for hypernymy, we sample 50 diagnostic tests from LM DIAGNOSTIC EXTENDED. We observe that in 10% of the examples, the model predicts the hyponym word itself (e.g., 'A yacht is a yacht.'). In 14% of examples, the model prediction is a valid hypernym that is not included in the Fischler categories. In 30% of diagnostic tests, BERT predicts a generic hypernym, often a part-of-speech (e.g., 'An imaret is a noun.') and in a further 12% BERT predicts a subword fragment of the hyponym as a hypernym, but this prediction is incorrect (e.g., 'A penknife is a pen.') We speculate that hypernyms often do occur in such patterns in the training data (for example, a *steamboat* is a *boat*), making such tests particularly difficult for BERT.[11] Finally, for 34% of the predictions the source of error is unknown; however, for 17.6% of these tests BERT defaults to predicting 'horse' and for 11.8% BERT predicts 'dog', suggesting that BERT may be assigning a higher prior to certain tokens when the prompt is unfamiliar. Table 3 further displays BERT predictions in the closed vocabulary setting. Surprisingly, we observe that BERT identifies hypernyms incorrectly in plural probes, even for frequently occurring hyponyms such as *'car'*, predicting *'cars are trees'*.

3.4 Frequency and Memorization Effects

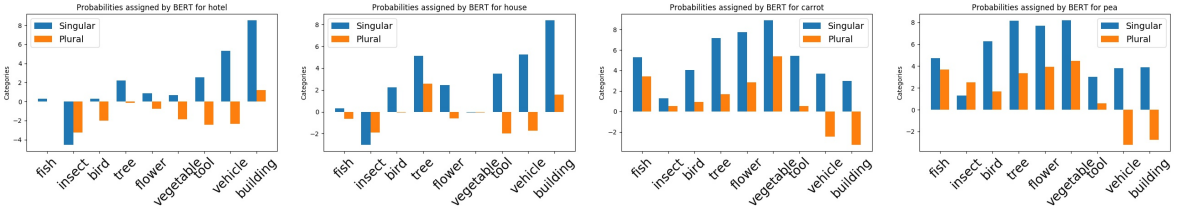
When does BERT fail to recognize hypernyms in the zero-shot probe setting? What role does term frequency play in this ability? We investigate two hypothesized failure modes. (1) Rare hyponym: How does BERT probe performance vary with term frequency? To examine this, we consider the frequency statistics of each hyponym in the LM DIAGNOSTIC EXTENDED diagnostic, and examine those where the hypernym relation is correctly identified by BERT. We observe that correctly recognized hyponyms tend to be significantly more frequent than unrecognized ones, occurring on average 5098.15 times in Wikipedia, compared with

[10] However, the open vocabulary setting of Ettinger (2020) suffers from the limitation that since there are many correct hypernyms for any target word, models may be unfairly penalized in this setting for predicting a hypernym not present in the diagnostic. For this reason, we further consider the closed vocabulary setting (Singular, Plural, Contextual and Paired Singular-Plural in Table 2), where we examine probabilities assigned by the PLM to the nine hypernym categories defined in Battig and Montague (1969).

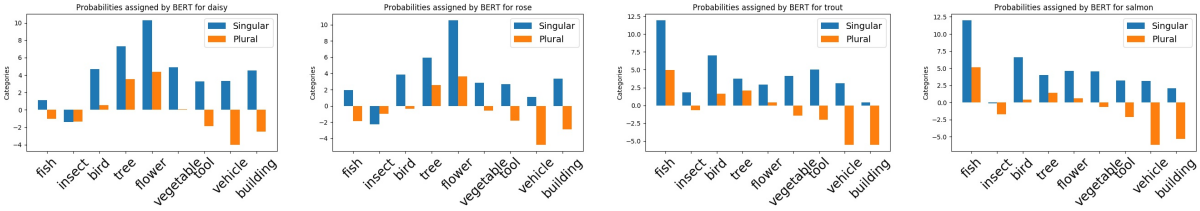
[11] Headed noun-noun compounds in English are likely to be right-headed (Williams, 1981).



(a) Log probabilities assigned by BERT to each category for singular and plural queries, for insect and vehicle hyponyms.



(b) Log probabilities assigned by BERT to each category for singular and plural queries, for building and vegetable hyponyms.



(c) Log probabilities assigned by BERT to each category for singular and plural queries, for flower and fish hyponyms.

Figure 2: Category-wise log probability predicted by BERT for singular and plural probes.

4359.55 times for unrecognized hyponyms.[12] (2) Pattern Matching: To examine this, we extract co-occurrence patterns between hyponym and hypernym for all pairs in LM DIAGNOSTIC EXTENDED. Of all the hyponym-hypernym pairs that are known to have occurred in the template *"[hyponym] is a [hypernym]"* on Wikipedia, we can predict the hypernymy relation correctly at 78.34%, considerably higher than the average performance on the diagnostic. These results suggest BERT may be acting as a sophisticated n-gram index, and be strong at retrieving facts it has explicitly seen before in the training data.

3.5 Singular and Plural Probes

What happens when a query is posed to BERT with plural number instead of singular? Figure 2 illustrates the probabilities assigned by BERT to each category for both singular and plural probes in the LM DIAGNOSTIC. We observe that in all cases, the correct answer is predicted with greater confidence when the probe is singular. We next analyze errors in model predictions in singular vs. plural probes. We find that overall 7.4% of tests are predicted correctly only in plural form, 30.9% only in singular form, *only 36.63% in both singular and plural forms*, and 25% in neither.

[12] We conduct a Shapiro-Wilk test for normality, allowing us to reject the null hypothesis of the frequency distributions being normal. We thus perform a Kruskal-Wallis non-parametric significance test, and find that recognized hyponyms tend to be statistically significantly more frequent ($p<0.05$).

Dataset	Seen Hypernyms	Unseen Hypernyms
Majority	50.00	50.00
Static Partial	56.32 ± 1.56	56.09 ± 4.21
Static	62.46 ± 3.99	58.15 ± 4.24
BERT Partial	48.36 ± 3.1	47.48 ± 4.98
BERT Context	**92.81** ± 0.81	**58.48** ± 1.74

Table 4: BERT results on hypernymy detection in SEEN and UNSEEN probing settings. Static is the summary of the best performance across word2vec, FastText and GloVe representations. The partial baseline for each representation, is the performance of a probing classifier trained only on the representation of the hypernym.

4 Paradigmatic Generalization

Representations: We use the following representations in the encode-embed-predict architecture described in §2.2.1: For static representation baselines, we use word2vec, GloVe and FastText, and for our contextualized representation we study BERT. Detailed descriptions of the architectures can be found in §3.2

Baselines: (1) Majority baseline: Performance of classifier that always predicts the majority class in the test set. (2) Partial: Partial-input baselines have revealed biases in Natural Language Inference (Tsuchiya, 2018; Gururangan et al., 2018; Poliak et al., 2018) and Question-Answering (Kaushik and Lipton, 2018) datasets. Levy et al. (2015) discuss the propensity of classifiers to rely on 'protoypical hypernyms' in hypernymy detection datasets, and not to solve the detection task. To control for potential dataset biases caused by the selection of items in the study, for each model we train a partial counterpart baseline, which is only provided the hypernym as input. If the dataset is unbiased in this aspect, partial baselines should achieve similar performance to a random classifier.

Results: Table 4 reports performance on SEEN and UNSEEN settings in our probing task. All experiments are done with 3-fold cross validation. We observe that all partial input baselines achieve near-random performance. Further, we observe that in the UNSEEN setting, probing classifier performance decreases considerably, indicating a lack of a systematic hypernymy function in BERT representations discoverable by the probing classifier. Thus, we determine that this class of probes does not generalize paradigmatically. Notably, we observe that a majority of the errors made by the probing classifiers is falsely detecting pairs as hypernyms, accounting for 79.4% of errors. Additionally, we observe that probing task design can considerably affect the conclusions drawn about whether a representation encodes any given property, emphasizing a need for careful consideration of design choices.

5 Related Work

There has been considerable interest in probing the capabilities of PLMs (Rogers et al., 2020). Much recent work focuses on the grammatical and syntactic capabilities of BERT (Hewitt and Manning, 2019; Liu et al., 2019; Swayamdipta et al., 2019; Goldberg, 2019; Wolf, 2019; Coenen et al., 2019; Tenney et al., 2019; Warstadt et al., 2019; Kim et al., 2019). In contrast, our focus is on probing studies that aim to uncover "knowledge" in BERT. There have been several such studies: Forbes et al. (2019) study physical commonsense encoded in BERT. Da and Kasai (2019) probe BERT for its understanding of object attributes, finding that it learns physical concrete norms (*is made of wood*) better than abstract ones (*is strong*). Wallace et al. (2019) find a 'surprising degree of numeracy' is present in contextualized word representations. Talmor et al. (2019) probe BERT for capabilities at particular types of symbolic reasoning, such as comparison, conjunction and composition.

Our work focuses specifically on the validity of conclusions drawn from such probing studies that aim to discover knowledge in BERT, using the setting of Ettinger et al. (2018) as a case-study. We further distinguish between the instrumentative and agentive perspectives on probing. For example, there has been considerable research attention focused on querying language models for their encoded information (Petroni et al., 2019; Jiang et al., 2020; Bosselut et al., 2019), which we consider as an instrumentative effort using PLMs as a tool. Our focus in this work is instead on agentive studies, and our conclusion is that the probes we study should not be used to reveal evidence of some systematic knowledge or competence in PLMs— although PLMs can still be utilized as tools to *extract* such knowledge from text.

Closest to our work, Kassner and Schütze (2020) find that PLMs do not differentiate between negated and non-negated statements. Negation is a notoriously hard phenomenon for neural NLP models (Morante and Sporleder, 2012; Fancellu et al., 2016; Naik et al., 2018); our work demonstrates that even affirmative factual knowledge that can be extracted from BERT does not systematically generalize. Our work is also closely related to recent challenge set construction efforts, which aim to serve as sanity checks on the knowledge and commonsense capabilities of models (Marelli et al., 2014; Naik et al., 2018; Glockner et al., 2018; Ribeiro et al., 2020). For example, McCoy et al. (2019) show that BERT finetuned for the natural language inference task, relies heavily on shallow heurestics instead of acquiring adequate commonsense knowledge. Our work is complementary, demonstrating through a simple consistency task that BERT's capabilities, as discovered through probes, may not correspond to some systematic general ability.

Our work examines, in particular, hypernymy knowledge encoded in BERT representations. The identification of hypernyms is studied extensively in cognitive science and philosophy. Some prominent theories include Rosch's category theory (Rosch and Lloyd, 1978) and Tversky's category resemblance approach (Tversky, 1977). This work

does not account for either of these interpretations of hypernymy, but instead relies on prior cognitive studies on category norms (Fischler et al., 1983; Battig and Montague, 1969) and relations defined with these super-ordinate categories in WordNet (Fellbaum, 1998b; Oltramari et al.). Additionally, our work ties into the rich history on modeling hypernymy in NLP systems (Lin, 1998; Weeds and Weir, 2003; Baroni et al., 2012; Rimell, 2014; Roller et al., 2014; Weeds et al., 2014; Shwartz et al., 2015; Vulić and Mrkšić, 2018) and evaluating distributional semantic models on their ability to represent it (Baroni and Lenci, 2011; Santus et al., 2015, 2016; Necsulescu, 2011; Vyas and Carpuat, 2017).

6 Discussion and Summary

We briefly discuss our findings and offer some guiding principles for future work.

Frequency and Memorization Effects: We find that BERT is particularly vulnerable to low-frequency phenomena in the training data, and succeeds at examples in the probe which have explicitly occurred in the training data. We speculate based on this evidence that BERT may just be memorizing the vast amount of training data it has been exposed to, rather than performing any kind of deeper reasoning.

Caution with cloze-style probes: BERT's Masked-LM format lends itself easily to cloze-style probes, which consider filling in a missing token correctly as evidence of PLM knowledge. Despite the accessibility of this format to investigate the behavior of PLMs, we speculate that, by design, the model is expected to fill in tokens whose context matches the provided template. The designer of the probing task may include templates to extract knowledge based on their intuitions, which (1) may or may not be the right template to extract the targeted kind of knowledge, (2) may provide enough inductive bias that it is unclear if the model understands the relation or understands how to match a particular template (which has been chosen so well based on the practitioners knowledge that it mimics the model actually understanding the deeper phenomena). We speculate that data-driven methods (Jiang et al., 2020; Bouraoui et al., 2019) can be designed to mitigate (2), but will exacerbate (1).

Dual Perspectives on PLMs: In this work, we characterize two perspectives on uncovering knowledge in PLMs: instrumentative and agent-based. We emphasize that while systematicity is a necessary requirement for agent-based analysis, as ideally we would like AI agents to reason like humans do, it is not necessary from an instrumentative perspective if the representations offer utility for a downstream task.

Implications for future work: In this work, we provide an investigation of current approaches to probing contextualized representations. Our tests for systematic generalization present a clearer picture of the conclusions that can be drawn from probing studies. We find that 'knowledge' discovered by standard probes does not serve to illuminate a systematic, general competence in the underlying PLMs. We suggest that future studies carefully evaluate the generalizability of their methods, and always be accompanied by consistency checks and controls to ensure that claims based on model behavior are made as reliable as possible.

Acknowledgements

We would like to thank Ali Emami, Ian Porada, Yanai Elazar, Alessandro Sordoni, Aakanksha Naik, Maria Ryskina and Shruti Rijhwani, for invaluable discussion related to the paper. We also thank the anonymous reviewers for their valuable suggestions. The last author is supported by the Canada CIFAR AI Chair program.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

Elena Akhmatova and Mark Dras. 2009. Using hypernymy acquisition to tackle (part of) textual entailment. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer)*, pages 52–60, Suntec, Singapore. Association for Computational Linguistics.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 23–32. Association for Computational Linguistics.

Marco Baroni and Alessandro Lenci. 2011. How we blessed distributional semantic evaluation. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10. Association for Computational Linguistics.

William F Battig and William E Montague. 1969. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental Psychology*, 80(3p2):1.

Or Biran and Kathleen McKeown. 2013. Classifying taxonomic relations between pairs of Wikipedia articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 788–794, Nagoya, Japan. Asian Federation of Natural Language Processing.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2019. Inducing relational knowledge from bert.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of bert.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Jeff Da and Jungo Kasai. 2019. Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 1–12, Hong Kong, China. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 495–504, Berlin, Germany. Association for Computational Linguistics.

Christiane Fellbaum. 1998a. Towards a representation of idioms in WordNet. In *Usage of WordNet in Natural Language Processing Systems*.

Christiane Fellbaum. 1998b. *WordNet: An Electronic Lexical Database*. Bradford Books.

Ira Fischler, Paul A Bloom, Donald G Childers, Salim E Roucos, and Nathan W Perry Jr. 1983. Brain potentials related to stages of sentence verification. *Psychophysiology*, 20(4):400–409.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Roxana Girju, Manju Putcha, and Dan Moldovan. 2003. Discovery of manner relations and their applicability to question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 54–60, Sapporo, Japan. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743.

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, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner and Hinrich Schütze. 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, Online. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094.

Aman Madaan, Dheeraj Rajagopal, Yiming Yang, Abhilasha Ravichander, Eduard Hovy, and Shrimai Prabhumoye. 2020. Eigen: Event influence generation using pre-trained language models.

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-2014)*, pages 216–223, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 791–799. Association for Computational Linguistics.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in*

NLP, pages 27–35, Atlanta, Georgia. Association for Computational Linguistics.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Silvia Necsulescu. 2011. Automatic acquisition of possible contexts for low-frequent words. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011*, pages 121–126, Hissar, Bulgaria. Association for Computational Linguistics.

Alessandro Oltramari, Aldo Gangemi, Nicola Guarino, and Claudio Masolo. Restructuring wordnet's top-level: The ontoclean approach.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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 (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

John Prager, Jennifer Chu-Carroll, Eric W Brown, and Krzysztof Czuba. 2008. Question answering by predictive annotation. In *Advances in Open Domain Question Answering*, pages 307–347. Springer.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Laura Rimell. 2014. Distributional lexical entailment by topic coherence. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 511–519.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1025–1036, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Eleanor Rosch and Barbara Bloom Lloyd. 1978. Cognition and categorization.

Enrico Santus, Alessandro Lenci, Tin-Shing Chiu, Qin Lu, and Chu-Ren Huang. 2016. Nine features in a random forest to learn taxonomical semantic relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4557–4564, Portorož, Slovenia. European Language Resources Association (ELRA).

Enrico Santus, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: an evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, Beijing, China. Association for Computational Linguistics.

Ferdinand de Saussure. 1916. Course in general linguistics (trans. wade baskin). *London: Fontana/Collins*, page 74.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Vered Shwartz, Omer Levy, Ido Dagan, and Jacob Goldberger. 2015. Learning to exploit structured resources for lexical inference. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 175–184, Beijing, China. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in neural information processing systems*, pages 1297–1304.

Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 801–808, Sydney, Australia. Association for Computational Linguistics.

Swabha Swayamdipta, Matthew Peters, Brendan Roof, Chris Dyer, and Noah A Smith. 2019. Shallow syntax in deep water. *arXiv preprint arXiv:1908.11047*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics – on what language model pre-training captures.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Tu Vu and Vered Shwartz. 2018. Integrating multiplicative features into supervised distributional methods for lexical entailment. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 160–166, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Vulić and Nikola Mrkšić. 2018. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, New Orleans, Louisiana. Association for Computational Linguistics.

Yogarshi Vyas and Marine Carpuat. 2017. Detecting asymmetric semantic relations in context: A case-study on hypernymy detection. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 33–43.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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 Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259. Dublin City University and Association for Computational Linguistics.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 81–88.

Henry M Wellman and Susan A Gelman. 1992. Cognitive development: Foundational theories of core domains. *Annual review of psychology*, 43(1):337–375.

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, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Edwin Williams. 1981. On the notions" lexically related" and" head of a word". *Linguistic inquiry*, 12(2):245–274.

Thomas Wolf. 2019. Some additional experiments extending the tech report" assessing berts syntactic abilities" by yoav goldberg. Technical report.

Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, and Gerhard Weikum. 2013. Robust question answering over the web of linked data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1107–1116. ACM.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Fold	#Train	#Dev	#Test
1	6164	206	232
2	7936	92	112
3	5892	222	219

Table 5: Statistics of UNSEEN dataset to study paradigmatic generalization.

Fold	#Train	#Dev	#Test
1	6682	176	148
2	6556	182	144
3	6582	164	154

Table 6: Statistics of SEEN dataset to examine paradigmatic generalization.

A Datasets for Paradigmatic Generalization

Table 5 and Table 6 summarize the dataset statistics of the unseen and seen datasets respectively. We perform 3-fold cross validation.