# Seeking Clozure: Robust Hypernym Extraction from BERT with Anchored Prompts

**Chunhua Liu** **Trevor Cohn**[*] **Lea Frermann**
School of Computing and Information Systems
The University of Melbourne
chunhua@student.unimelb.edu.au
{tcohn,lfrermann}@unimelb.edu.au

## Abstract

The automatic extraction of hypernym knowledge from large language models like BERT is an open problem, and it is unclear whether methods fail due to a lack of knowledge in the model or shortcomings of the extraction methods. In particular, methods fail on challenging cases which include rare or abstract concepts, and perform inconsistently under paraphrased prompts. In this study, we revisit the long line of work on pattern-based hypernym extraction, and use it as a diagnostic tool to thoroughly examine the hypernomy knowledge encoded in BERT and the limitations of hypernym extraction methods. We propose to construct prompts from established pattern structures: definitional (*X is a Y*); lexico-syntactic (*Y such as X*); and their anchored versions (*Y such as X or Z*). We devise an automatic method for anchor prediction, and compare different patterns in: (i) their effectiveness for hypernym retrieval from BERT across six English data sets; (ii) on challenge sets of rare and abstract concepts; and (iii) on consistency under paraphrasing. We show that anchoring is particularly useful for abstract concepts and in enhancing consistency across paraphrases, demonstrating how established methods in the field can inform prompt engineering.[1]

## 1 Introduction

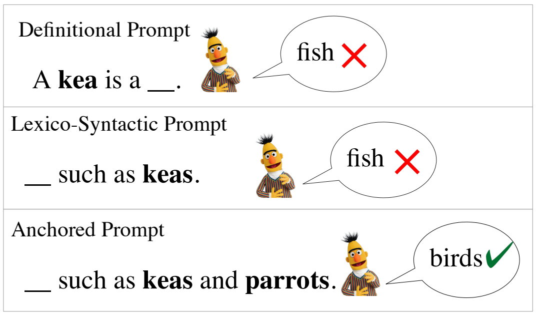

Figure 1: Example prompts for hypernym prediction, derived from established pattern structures.

Semantic relations play a central role in knowledge representation (Miller, 1995) and taxonomy construction (Snow et al., 2006; Navigli et al., 2011). As the backbone of semantic relations, hyponymy/hypernymy relations express a hierarchical relation between a specific concept (the hyponym; e.g., dog) and a general one (the hypernym; e.g., mammal), and form the foundation of human concept understanding (Yu et al., 2015) and relation reasoning (Lyons, 1977; Green et al., 2002). Given its fundamental role, the automatic extraction of hypernym knowledge from large texts (Hearst, 1992; Roller et al., 2018) or pre-trained language models (PLMs) (Takeoka et al., 2021; Jain and Espinosa Anke, 2022), and its injection into NLP methods are active areas of research (Peters et al., 2019).

The unsupervised extraction of hypernyms from PLMs by prompting has attracted recent attention, e.g., using patterns like *A dog is a type of [MASK]* and retrieving the most likely filler words from the model (Ettinger, 2020; Weir et al., 2020; Jain and Espinosa Anke, 2022). Results were mixed: while PLMs can reliably predict hypernyms of concrete and frequent hyponyms (Ettinger, 2020; Weir et al., 2020), experiments on more challenging data sets show a quick deterioration in the face of rare concepts (Schick and Schütze, 2019), and a lack of response consistency across paraphrased prompts (Ravichander et al., 2020; Elazar et al., 2021). How to alleviate these issues and extract more reliable hypernyms from PLMs remain open questions.

In this paper, we draw connections between prompting for hypernyms and pattern-based hypernym extraction (Hearst, 1992; Snow et al., 2004) (see Figure 1 and Table 1). We systematically investigate the utility of different styles of patterns as BERT prompts, and use them as diagnostic tools

[*] Now at Google DeepMind.

[1] Code and test sets are available at https://github.com/ChunhuaLiu596/AnchoredPrompts

to better understand the conditions under which probing for hypernyms is effective and consistent.

Pattern-based hypernym extraction from raw text has a long history, starting from Hearst (1992)'s seminal work which promotes lexico-syntactic patterns (*Y such as X*)[2] as more effective than definitional patterns (*X is a type of Y*). Follow-up work (Hovy et al., 2009) incorporated a co-hyponym, a concept that shares a hypernym with X, into the pattern (*Y such as X and Z*) to provide additional context signals. Figure 1 illustrates this, where the anchor *parrot* provides additional information to facilitate the prediction of the correct hypernym of *kea*. This method of 'anchoring' has been shown to improve the quality of automatically extracted hypernym knowledge. We apply these established patterns from the hypernym extraction literature in the context of language model prompting, and systematically study the existence and gaps of hyponym/hypernym knowledge in BERT. We conduct experiments on six English data sets and address three questions:

*How to effectively construct anchored prompts?* We devise a scalable method to automatically retrieve high-quality anchors (co-hyponyms) to construct anchored prompts. Anchors are mined from PLMs with established co-hyponym patterns (e.g., *such as X and* ___) and evaluated with WordNet (Miller, 1995).

*How do different pattern structures compare as prompts under different data conditions?* We ground our prompts in hypernym patterns from which have been successfully used to mine hypernyms from raw corpora, and investigate their effectiveness for zero-shot PLM hypernym retrieval. We find strong, consistent benefits of anchored prompts, particularly for rare or abstract concepts.

*Robust extraction of hypernym knowledge.* Much recent work has shown that PLM prompting results are brittle under prompt paraphrases, calling into question whether prompting surfaces robust knowledge encoded in the PLMs or rather superficial associations. We compare the robustness of different patterns under paraphrasing, and find, again, a benefit of anchored prompts for retrieving more consistently correct hypernyms.

In summary, we contribute to the on-going research on hypernym extraction by unifying the long-standing work of pattern-based and prompt-based approaches, demonstrating that anchoring prompts can unlock a wealth of hidden knowledge within BERT, and providing a framework of automatic construction of anchoring prompts.

| | | | |
|---|---|---|---|
| DFP | A(n) X is a Y.<br>A(n) X is a type of Y.<br>A(n) X is a kind Y. | DFP$^{+A}$ | A(n) X or Z is a Y.<br>A(n) X or Z is a type of Y.<br>A(n) X or Z is a kind Y. |
| LSP | Y such as X.<br>Y, including X.<br>Y, especially X.<br>X or other Y.<br>X and other Y.<br>such Y as X. | LSP$^{+A}$ | Y such as X and Z.<br>Y, including X and Z.<br>Y, especially X and Z.<br>X, Z or other Y.<br>X, Z and other Y.<br>such Y as X and Z. |

Table 1: Four types of pattern structures: definitional patterns (DFP; top) and lexico-syntactic patterns (LSP; bottom); and their anchored versions: DFP$^{+A}$ and LSP$^{+A}$ (right).

## 2 Background

We introduce the two approaches for hypernym extraction on which we build in this paper: pattern-based (§ 2.1) and prompting PLMs (§ 2.2).

### 2.1 Pattern-based Hypernym Extraction

The pattern-based approach applies hyponym-hypernym patterns to large corpora to extract hypernyms. Two widely-used pattern structures have been identified: lexico-syntactic and definitional.

### 2.1.1 Lexico-Syntactic Patterns (LSP)

Lexico-syntactic patterns (LSP; Table 1 bottom left) were first introduced by Hearst (1992) and have since been used to mine hyponym-hypernym pairs or build ontologies from large corpora (Pasca, 2004; Pantel and Pennacchiotti, 2006; Etzioni et al., 2005; Roller et al., 2018). The six LSP (1) all indicate the hyponym-hypernym relation with explicit signals (e.g., *such as, especially*), (2) frequently occur in text, and (3) are applicable to nouns or noun-phrases.

**Anchored LSP (LSP$^{+A}$)** Hovy et al. (2009) proposed an 'anchored' version of LSP to mine hypernyms (LSP$^{+A}$)[3] which uses patterns like *Y such as X and Z*, where Z is an anchor which reduces ambiguity and assists the extraction of Y (Table 1, bottom right). A similar idea of using anchors to improve hypernym classifiers is used in Snow et al. (2004) and Bernier-Colborne and Barrière (2018). LSP$^{+A}$ has been shown to be effective at extracting reliable hypernyms from text corpora, however,

[2] We use $Y$ to denote hypernyms, $X$ for hyponyms and $Z$ for the co-hyponym of $X$.

[3] LSP$^{+A}$ is referred to as DAP$^{-1}$ in the original paper.

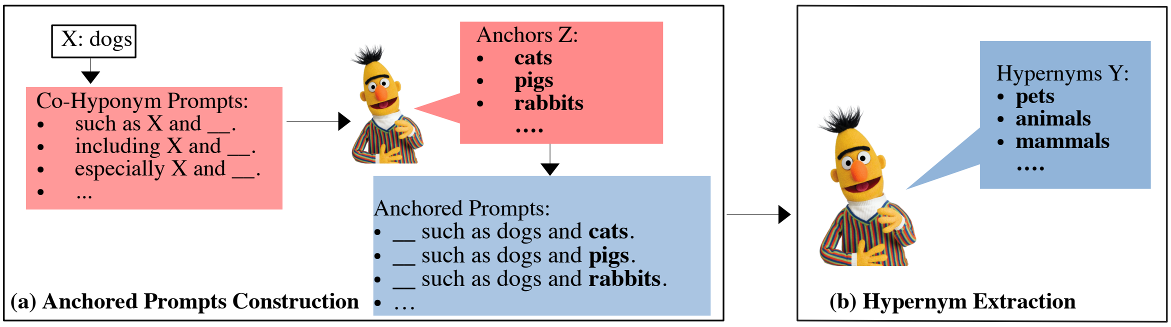

Figure 2: The workflow of constructing anchored prompts (a) and extracting hypernyms from PLMs (b).

like all pattern-based approaches, it suffers from low recall because it needs X, Y and Z to co-occur. The sparsity issue can be potentially remedied by using embeddings from PLMs to represent X and Z when used as prompts. However, this hasn't been studied in the context of extracting knowledge from PLMs. Inspired by this line of work and PLM prompting, we use LSP$^{+A}$ to mine hypernyms from PLMs and examine the benefit of anchors.

### 2.1.2 Definitional Patterns (DFP)

In contrast to LSP that conveys the hypernym relation implicitly, definitional Patterns (DFP; Table 1 top left) explicitly define an *Is-A* relation between X and Y (Lyons, 1977). A common use of DFP is to mine sentences for definition extraction (Borg et al., 2009; Navigli et al., 2010) or ontology/dictionary building (Muresan and Klavans, 2002). Recently, DFP has been widely used in prompting studies (Schick and Schütze, 2020; Ettinger, 2020; Ravichander et al., 2020; Hanna and Mareček, 2021) to probe hypernym knowledge in PLMs.

**Anchored DFP (DFP$^{+A}$)** Analogous to LSP$^{+A}$, we augment DFP with anchors for disambiguation (Table 1 top right). To the best of our knowledge, Hanna and Mareček (2021) is the only work which uses anchored definitional patterns to prompt PLMs for hypernyms, described in more detail below.

## 2.2 Prompting-based Hypernym Extraction

With recent advances in PLMs, increasingly rich knowledge is captured language models. A stream of research aims at automatically extracting this knowledge, e.g., by probing PLMs for hypernym knowledge (Ettinger, 2020; Weir et al., 2020; Peng et al., 2022). Hanna and Mareček (2021) examined the effects of single hypernym patterns (e.g., *'X is a Y'*, *'A Y such as X'*) on prompting PLMs and showed that performance varies with patterns. Similarly, Ravichander et al. (2020) found that PLMs fail to retrieve consistent knowledge over prompts paraphrased with singular vs plural hyponyms.

Most previous work on prompting was conducted under relatively simple conditions with one pattern structure and a single data set. We systematically investigate the effects of well-established patterns (LSP/LSP$^{+A}$ and DFP/DFP$^{+A}$) on extracting hypernyms across six widely-used datasets and paint a more nuanced picture of hypernym knowledge in BERT by explicitly studying the challenging cases of rare or abstract concepts.

# 3 Anchored Prompts

We now introduce our framework of extracting hypernyms from a PLM by constructing sets of prompts given a hyponym X and a pattern type $\in$ {DFP, DFP$^{+A}$, LSP, LSP$^{+A}$ }. We illustrate the workflow in Figure 2, with LSP$^{+A}$ as an example.

**Prompt Construction** For each pattern type, we construct a set of prompts by instantiating each of its assigned patterns (cells in Table 1) with a concept in positions X and Z, and a [MASK] token in position Y. For DFP and LSP we can construct prompt sets directly given a hyponym X of interest. To construct prompts for LSP$^{+A}$ and DFP$^{+A}$ we need to additionally provide meaningful anchors Z. We next describe a way to effectively mine such anchors from language models (see Figure 2 (a)).

**Anchor Extraction** Given X, we use BERT to automatically extract a set of anchors, i.e., concepts Z that share a hypernym with X. To acquire such anchors, we again adopt a set of established lexico-syntactic patterns that indicate the fact that X and Z share a common hypernym (Hearst, 1992; Snow et al., 2004; Etzioni et al., 2005). Table 2 presents the full list of patterns we used to mine

| | |
|---|---|
| such as X and Z. | including X and Z. |
| such as X or Z. | including X or Z. |
| such as X, Z, | including X, Z, |
| especially X and Z. | X, Z or other |
| especially X or Z. | X, Z and other |
| especially X, Z, | |

Table 2: Co-hyponym patterns for anchor extraction, adapted from Hearst (1992).

anchors. Each pattern is converted into a prompt by filling in X and replacing Z with a [MASK] token, resulting in a set of co-hyponyms prompts $\mathcal{C}$. We retrieve the 10 most likely filler words according to language model probability for each pattern $\mathcal{C}_i \in \mathcal{C}$. We score candidates $z$ by their average probability across the patterns that contained $z$ among the top 10 fillers:

$$s_{LM}(z|x, \mathcal{C}) = \frac{1}{|\mathcal{C}_z|} \sum_{i=1}^{|\mathcal{C}|} P_{LM}(z|x, \mathcal{C}_i), \quad (1)$$

where $P_{LM}(z|x, \mathcal{C})$ is the probability of $z$ in the i$^{th}$ pattern instantiated with $x$ and $|\mathcal{C}_z|$ is the number of patterns that predicted $z$. We finally keep the $M$ highest scoring concepts as anchors, and instantiate $M$ copies of $\texttt{LSP}^{+\texttt{A}}$ and $\texttt{DFP}^{+\texttt{A}}$ with the different anchors, respectively.

**Hypernym Extraction** Being able to construct sets of prompts for vanilla ($\mathcal{P}_{\texttt{DFP}}, \mathcal{P}_{\texttt{LSP}}$) and anchored prompts ($\mathcal{P}_{\texttt{DFP}^{+\texttt{A}}}, \mathcal{P}_{\texttt{LSP}^{+\texttt{A}}}$), we are now in a position to prompt PLMs for hypernyms. Separately for each prompt set $\mathcal{P}$,[4] we score hypernym candidates $y$ by their average probability across patterns $\mathcal{P}_i \in \mathcal{P}$:

$$s_{LM}(y|x, \mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \log P_{LM}(y|x, \mathcal{P}_i), \quad (2)$$

where $\mathcal{P} = \{\mathcal{P}_{\texttt{LSP}}, \mathcal{P}_{\texttt{DFP}}, \mathcal{P}_{\texttt{LSP}^{+\texttt{A}}}, \mathcal{P}_{\texttt{DFP}^{+\texttt{A}}}\}$. The hypernyms ranked by $s_{LM}(y|x, \mathcal{P})$ and the top $K$ are retained as hypernym candidates.

## 4 Experimental Setup

**Datasets** We conduct experiments on six English datasets. CLSB (Devereux et al., 2014) and DIAG (Ravichander et al., 2020) have been recently used to probe for hypernym knowledge in PLMs (Devlin et al., 2019). The remaining four data sets are widely-used test sets for hypernym extraction more generally (Shwartz et al., 2017; Roller et al., 2018), namely BLESS (Baroni and Lenci, 2011), EVAL (Santus et al., 2015), LEDS (Baroni et al., 2012), and SHWARTZ (Shwartz et al., 2017). We only consider NOUN-NOUN hyponym-hypernym pairs from the datasets. Dataset statistics are reported in Table 9 in the Appendix. Data sets vary widely in terms of their corpus size, the ratio of abstractness and concreteness, concept frequency and their construction methods, and hence underlying knowledge sources. While most data sets are based on WordNet, SHWARTZ builds on a wider set of resources, including ConceptNet and Wikipedia and hence includes more obscure concepts. EVAL stands out with a relatively high proportion of abstract concepts, unlike the other data sets which are predominantly concrete. Section 6 explores performance using these data conditions.

**Model** All our experiments are based on BERT-large-uncased (Devlin et al., 2019) from Huggingface[5] and use a zero-shot approach to probe the model. To allow for comparability of results across data sets, we adopt an *open vocabulary* approach throughout, considering the whole BERT vocabulary as hypernym candidates.[6] We remove test instances where the hypernym is not in the BERT vocabulary.[7] We set the number of anchors in anchored prompts to $M = 5$.[8]

**Evaluation Metrics** Following previous work (Petroni et al., 2019; Qin and Eisner, 2021), we retain the $K$=10 hypernym candidates and report Precision at 10 (P@10) as the extent to which correct hypernyms are included in the top 10 model predictions ranked by Equation 2. We also report mean reciprocal rank (MRR) of the true label. We evaluate model predictions at the *concept level*, normalizing predictions into their canonical form, i.e., accepting any inflection of the correct hypernym,[9] and exclude punctuation, stop words, numbers and the hyponym $x$ from the predictions.

[4] We drop subscripts to avoid clutter.

[5] `https://huggingface.co/bert-large-uncased`

[6] Prior work (Ravichander et al., 2020) adopted a *closed-vocabulary* approach, limiting the set of candidate $y$ to hypernyms in a particular data set.

[7] Note that there is no such restriction on hyponyms so that results in § 6.2 are not biased.

[8] This number was optimized on BLESS.

[9] We used pyinflect 0.5.1.

| Dataset | MRR | P@1 | P@5 | P@10 |
|---|---|---|---|---|
| BLESS | 73.9 | 66.0 | 86.6 | 89.6 |
| DIAG | 34.9 | 28.6 | 43.8 | 48.8 |
| CLSB | 60.3 | 51.2 | 73.2 | 77.7 |
| SHWARTZ | 23.7 | 16.8 | 33.1 | 39.8 |
| EVAL | 33.6 | 26.1 | 44.1 | 49.4 |
| LEDS | 45.8 | 35.7 | 59.7 | 66.3 |

Table 3: Anchor evaluation results, where predicted anchors $z$ for a concept $x$ are validated by checking whether $x$ and $z$ share a hypernym in WordNet.

We measure the significance of differences with paired t-tests at $p<0.05$ after Holm-Bonferroni correction for multiple comparisons to adjust for comparisons across six data sets (Dror et al., 2017).

**Analyses** In addition to the main results, we aim to understand underlying factors that might affect the performance of hypernym extraction. We analyse the performance of pattern types on different types of concepts. We distinguish sets of hyponyms and hypernyms in terms of their frequency and abstractness and test consistency of predictions across prompt paraphrases.

## 5 Anchor Validation

*How accurate are the automatically mined anchors?* We qualitatively and quantitatively inspect retrieved anchor concepts. We use WordNet for this purpose, and follow Schick and Schütze (2020) to consider a candidate $z$ to be a valid anchor of $x$ if they share a common ancestor, within two levels above $x$ and four levels above $z$. We exclude hyponyms that are not in WordNet in this analysis.

Table 3 reports the results across six datasets. For three of the data sets (BLESS, CLSB, LEDS), a correct anchor is predicted as top 1 result more than 33% of the time, and contained among the top 10 predictions we consider close to 70% of the time. The other data sets are overall challenging due to diversity and/or low frequency of concepts.

Qualitative inspection reveals that retrieved anchors that are not WordNet siblings according to our definition above are often reasonable, see Table 4. As we shall see in Section 6 the utility of anchors does not seem to hinge on them being actual co-hyponyms, and that the topically related anchors as produced by our method effectively improve hypernym extraction.

| $x$ | Top 5 predicted anchors ($\mathcal{Z}$) |
|---|---|
| car | **truck**, **motorcycle**, **boat**, yes, **bike** |
| apple | **grape**, **pear**, nuts, vegetable, **date** |
| train | **bus**, plane, **car**, **tram**, **truck** |
| corn | bean, potato, **barley**, **wheat**, pea |
| panzer | **tank**, infantry, gun, artillery, panther |
| motel | hotel, yes, sure, restaurant, actually |
| daisy | rose, yes, lavender, rush, fern |
| murre | dog, bird, fox, crow, rabbit |
| trireme | warship, frigate, ship, ferry, battleship |

Table 4: Examples of mined anchors ($\mathcal{Z}$) for hyponyms that share $\geq 1$ (top) or zero (bottom) co-hyponyms with WordNet. Anchors confirmed in WordNet in bold.

## 6 Hypernym Evaluation

We first examine the effectiveness of `LSP` vs `DFP` and the added value of anchoring on our six data sets overall (§ 6.1). Afterwards, we inspect specifically rare (§ 6.2) and abstract (§ 6.3) concepts as well as the well-known issue of inconsistency of responses in the face of prompt paraphrases (§ 6.4), explore different patterns in these contexts and end with an error analysis (§ 6.5).

### 6.1 Main Results

Table 5 presents the main results. Performance over datasets varies widely, with SCHWARTZ standing out with particularly low performance. SHWARTZ is dominated by proper noun hyponyms (e.g., city/person names), and includes a very broad range of hypernyms (1.1K). Performance on the other data sets are more comparable.

*Do `LSP` and `DFP` differ?* Comparing row one (`DFP`) and three (`LSP`) in Table 5, we see no consistent trend. While performance is often comparable, on BLESS `LSP` outperforms `DFP`. The reverse is true for EVAL. BLESS contains frequent and largely unambiguous hyponyms which are presumably more frequently discussed in natural patterns as comprised by `LSP`. EVAL is dominated by ambiguous and abstract concepts, which are perhaps more commonly described by formal, definition-style language.

*Do anchors help retrieve more accurate hypernym knowledge?* Table 5 reveals a consistent improvement of adding anchors for `DFP` (row 1 vs. 2) but not for `LSP` (row 3 vs. 4): definitional patterns benefit from anchoring via co-hyponyms

| | BLESS | | DIAG | | CLSB | | SHWARTZ | | EVAL | | LEDS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | P@10 | MRR | P@10 | MRR | P@10 | MRR | P@10 | MRR | P@10 | MRR | P@10 |
| DFP | 23.6 | 42.4 | 42.6 | 66.8 | 39.8 | 67.5 | 6.3 | 12.8 | **24.0** | **46.7** | 32.6 | 60.1 |
| DFP$^{+A}$ | 25.7$^{+}$ | 47.2$^{+}$ | **45.5**$^{+}$ | **67.2**$^{+}$ | **42.3**$^{+}$ | **70.5**$^{+}$ | 5.9$^{+}$ | 13.6$^{+}$ | 22.1$^{+}$ | 43.3$^{+}$ | **35.7**$^{+}$ | **64.3**$^{+}$ |
| LSP | **27.1**$^{*}$ | **53.9**$^{*}$ | **45.5** | 66.1 | 40.8 | 68.2 | 6.4 | **15.2**$^{*}$ | 17.3$^{*}$ | 39.5$^{*}$ | 33.4 | 60.5 |
| LSP$^{+A}$ | 26.5 | 53.2 | 42.8$^{+}$ | 62.7 | 40.4 | 67.7 | **6.5** | 14.9 | 17.0 | 38.1 | 34.0 | 61.6 |

Table 5: Main results on six hypernym extraction datasets. Bold number indicates the highest score per data set and metric. $^{*}$ indicates significant difference of LSP vs. DFP; $^{+}$ indicates significant difference wrt. the non-anchored counterpart (i.e., LSP vs LSP$^{+A}$ and DFP vs. DFP$^{+A}$).

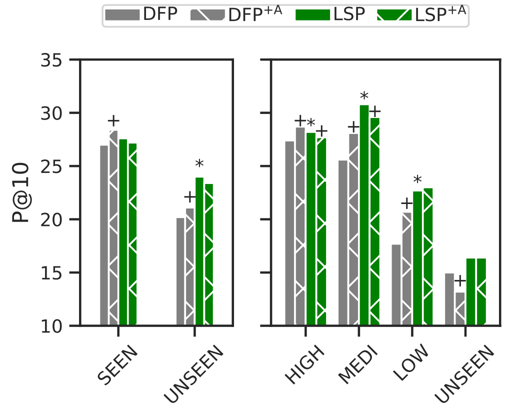

Figure 3: Performance of different pattern structures rare vs common hyponyms. Left: hyponyms seen in BERT vocabulary and not. Right: hyponyms frequency of different frequency bands estimated from large corpora. $^{+}$ and $^{*}$ as in Table 5.

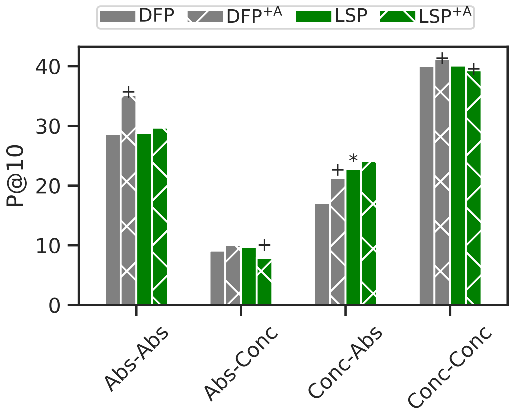

Figure 4: Performance of different pattern structures on: abstract hypo- and hypernym (Abs-Abs); abstract hypo- concrete hypernym (Abs-Conc); concrete hypo- abstract hypernym (Conc-Abs); and concrete hypo- and hypernym (Conc-Conc). $^{+}$ and $^{*}$ as in Table 5.

while lexico-syntactic patterns don't.[10]

Next, in §6.2-§6.4 we disentangle the main results, considering a range of conditions which have been identified as challenging in prior work, and examine whether different patterns and/or anchoring can improve hypernym retrieval from BERT in these contexts.

## 6.2 The Impact of Frequency

Previous work (Ravichander et al., 2020; Hanna and Mareček, 2021; Schick and Schütze, 2020) found that *BERT often fails to predict hypernyms for uncommon hyponyms.* Here, we examine whether incorporating anchors can alleviate this issue. This is driven by the intuition that humans often draw on surrounding context signals to help understand the relationship between concepts. For example, even if we are unfamiliar with the concept of *kea*, knowing an anchor like *parrot* can help us infer that *bird* is one of the hypernyms. We expect that anchors can provide more linking context to the hypernym and improve the hypernym extraction performance when the hyponyms are rare. To verify this, we look into two aspects that reflect frequency: (a) existence in the BERT vocabulary - hyponyms that are included as single-tokens are frequent; (b) frequency in large corpora. We obtain term frequency from WorldLex (Gimenes and New, 2016) and categorize frequency into four levels based on absolute count: High (> 100), Medium (10-100), Low (1-10), and Unseen (0). For this analysis, we aggregate instances from all datasets to increase statistical power.

Figure 3 presents experimental results. We find that rare hyponyms have lower performance in general, aligning with previous work (Ravichander et al., 2020; Hanna and Mareček, 2021). More in-

[10] We estimate the upper-bound of anchoring prompts with oracle anchors from WordNet, finding that better anchors can bring more benefits (see Table 10 in the Appendix).

| $x$ | DFP Predictions | DFP$^{+A}$ Predictions | Top 5 predicted anchors ($\mathcal{Z}$) |
|---|---|---|---|
| terebinth | stone, sculpture, rock | **tree**, plant, sculpture | fern, shell, plant, shrub, tree |
| dray | boat, machine, tool | **vehicle**, cart, wagon | wagon, tractor, cart, horse, yes |
| gannet | computer, net, network | **bird**, fish, dolphin | seal, dolphin, herr, whale, penguin |
| happiness | joy, life, pleasure | joy, feeling, **emotion** | love, joy, good, personal, maybe |
| principle | rule, law, concept | rule, law, **value** | practice, rule, procedure, guideline, value |
| snoopy | toy, pigeon, mouse | toy, puppet, **character** | peanut, snoop, batman, garfield, cartoon |

Table 6: Examples of rare (top) and abstract (bottom) hyponyms $x$, along with their predicted hypernyms from DFP and DFP$^{+A}$, and predicted anchors. Correct hypernyms are in **bold**.

| | Singular Probes | Plural Probes | BLESS | DIAG | CLSB | SHWARTZ | EVAL | LEDS |
|---|---|---|---|---|---|---|---|---|
| DFP | A(n) X is a(n) Y. | X are Y. | 2.7 | 4.5 | 3.5 | 0.4 | 1.7 | 4.8 |
| DFP$^{+A}$ | A(n) X or Z is a(n) Y. | X or Z are Y. | 0.2 | 2.3 | 0.3 | 0.0$^{+}$ | 0.1 | 0.6 |
| LSP | Y such as a(n) X. | Y such as X. | **51.2** | 46.0 | 60.9 | 4.4 | 26.2 | 40.6 |
| LSP$^{+A}$ | Y such as a(n) X or Z. | Y such as X or Z. | **51.2** | **51.6**$^{+}$ | **65.0**$^{+}$ | **10.4**$^{+}$ | **32.5**$^{+}$ | **52.5**$^{+}$ |

Table 7: Experimental results (P@10) on pairwise number consistency. X/Z in singular probes are instantiated as singular (e.g., car), and in plural probes as plural (e.g., cars). $^{+}$ as in Table 5.

| | BLESS | DIAG | CLSB | SHWARTZ | EVAL | LEDS |
|---|---|---|---|---|---|---|
| DFP | 21.9 | 42.5 | 44.7 | 4.6 | 23.7 | 34.3 |
| DFP$^{+A}$ | **31.7**$^{+}$ | **49.0** | **53.8**$^{+}$ | **8.3**$^{+}$ | **28.3**$^{+}$ | **42.2**$^{+}$ |
| LSP | 26.8 | 32.8 | 45.8 | 2.6 | 10.2 | 29.0 |
| LSP$^{+A}$ | **31.7**$^{+}$ | **39.9**$^{+}$ | **52.5**$^{+}$ | **4.7**$^{+}$ | **13.3**$^{+}$ | **36.1**$^{+}$ |

Table 8: Experimental results (P@10) on group consistency. $^{+}$ as in Table 5.

terestingly, unlike in the main results, LSP exhibits a significant advantage over DFP on unseen and low frequency hyponyms (solid bars in UNSEEN and LOW blocks in Figure 3). Moreover, on the same blocks, we see that incorporating anchors into DFP significantly improves the performance on low frequent hyponyms (solid gray vs dashed gray). This confirms our hypothesis that anchors are beneficial for uncommon hyponyms by guiding BERT to predict hypernyms (see examples in Table 6). This is of practical relevance as it demonstrates that anchored prompts help for uncommon hyponyms, which can inform hypernym extraction in domain-specific or low-resources situations.

### 6.3 The Impact of Concreteness

Previous work on distributional semantics has shown that abstract words have higher contextual variability and are more difficult to predict than concrete concepts (Naumann et al., 2018). Here, we examine specifically whether the degree of concept abstractness affects hypernym extraction accuracy, as well as the impact of different patterns and anchoring in this context. To obtain the concept concreteness level, we use the Brysbaert dataset (Brysbaert et al., 2014),[11] which covers abstractness ratings for 40K common English concepts. Each concept was scored by at least 25 human annotators on a scale from 1 (most abstract) to 5 (most concrete). We use the median score to represent the abstractness of each word and bin them into Abstract ($< 3$) and Concrete ($\geq 3$). We inspect all four possible combinations of {concrete, abstract} × {hypernym, hyponym}, and again aggregate instances across data sets.

Figure 4 shows that hypernyms of hyponyms at same abstraction levels (e.g., Conc–Conc) are predicted with higher accuracy than those under different levels (e.g., Abs-Conc). This result is intuitive as words in same abstraction level tend to co-occur more (Bhaskar et al., 2017; Frassinelli et al., 2017). Overall, concrete hyponym-hypernym pairs are predicted with higher accuracy than pairs involving an abstract concept, indicating that abstract knowledge is more difficult to retrieve from BERT. More interestingly, we find that DFP$^{+A}$ brings remarkable improvements on abstract hypernyms, effectively reducing the gap between abstract and concrete hypernyms. A closer look at abstract hypernyms that failed with DFP but succeed on anchored prompts reveals failure on abstract hypernyms such as {*emotion, organization, language, event*}. For example, for the prompt *excitement is a* ___ BERT predicts {*thrill, fear, rush*}. However, by incorporating anchors like *surprise* or *anxiety*, BERT predicts the

[11] We exclude hyponyms and hypernyms that are not in the Brysbaert dataset.

correct hypernym *emotion*. This finding is encouraging because it points to the weakness of using hyponyms alone to prompt PLMs for abstract hypernyms and can potentially inform future work on prompt design for retrieving specific types of knowledge (e.g., concrete or abstract) and building ontologies.

### 6.4 Consistency

Despite the success of prompting, a persistent challenge is an inconsistency of responses under slight rephrasing of the prompt (Elazar et al., 2021). In the context of hypernomy prediction, Ravichander et al. (2020) showed that compared to singular prompts (*a car is a* ___.), plural versions (*cars are* ___.) returned different and worse results. We study consistency more systematically by including different paraphrases, and exploring the utility of anchoring on the robustness of results. We investigate: (a) consistency across prompts paraphrased with singular and plural hyponyms; and (b) consistency over prompts paraphrased with pattern type instantiations (cells in Table 1). We only score the prediction for a test instance as correct, if it was correctly predicted by *all* prompt paraphrases.

**Pairwise Number Consistency** Following Ravichander et al. (2020), we construct pairwise probes for singular and plural hyponyms, obtaining one representative pair for each of our four pattern types as listed in Table 7 (left). The results in Table 7 show that consistency strongly correlates with the choice of patterns: DFP prompts (row 1) produce inconsistent results, while LSP (row 3) shows strong potential for retrieving consistent knowledge. One reason is ambiguity in the plural DFP: the prompt *Xs are [MASK]* tends to return verbs and adjectives as candidates (e.g., *carrots are {grown, eaten, orange}.*), as plausible completions. In contrast, LSP contexts are more specific. Moreover, the consistency improves significantly for all but one data set when incorporating the anchors into LSP.[12] This finding is important as it identifies a promising means of retrieving consistent knowledge from PLMs.

**Group Consistency** Our sets of pattern-type specific prompts suggest a natural, stricter consistency evaluation, namely to test whether BERT reliably predicts the same, true hypernym for all prompts associated with a pattern type (i.e, each of the cells of Table 1). Table 8 presents the results. What stands out in the table is that anchored prompts significantly improve group consistency, which aligns with our observation in the pairwise number consistency tests above. In summary, our results show that anchors, in particular LSP$^{+A}$, can help retrieve more robust and consistent hypernyms from PLMs. This is not only important for downstream tasks which rely on (automatic) high-quality hypernym knowledge, such as taxonomy creation, but could also inform strategies to probe BERT for genuine, systematic knowledge, rather than superficial associations.

### 6.5 Error Analysis

*When does anchoring hurt?* Beyond benefits from anchors, we also observed that incorporating anchors at times degrades performance. Closer inspection identified **sense ambiguity** as a prevalent reason, especially for polysemous hyponyms, which have multiple hypernyms of different senses (e.g., fan is a person or an appliance.) With anchors, BERT predictions are skewed to a specific sense as selected by the anchor, which can be different from the true hypernym. Another situation is noisy anchors, including generic and irrelevant anchors (e.g., actually), or topically related anchors that are not co-hyponyms (e.g., wood and lake).

*How do anchors improve consistency?* We analyse hypernyms that are not consistently predicted correctly without anchors but are correct with anchors. There are three reasons for the inconsistency: (a) overly generic predictions from non-anchored patterns, e.g., *Y, especially X* often produces hypernyms like *things* or *items*; (b) predictions of co-hyponyms instead of hypernyms without anchors (*e.g., a dog is a cat.*), which is especially common with pattern *A X is a Y*, for which 30% of its predictions contain co-hyponyms from WordNet; (c) hypernyms in the intermediate levels of the WordNet taxonomy (e.g., garment, jewelry, sweet) are less consistent for patterns without anchors, e.g., anchors improve consistency by 11% for hypernyms whose minimum taxonomy depth is 7.[13] This suggests that anchors can improve the consistency of mining new intermediate hypernyms from PLMs, aligning with prior work of using anchors to mine intermediate hypernyms from corpora (Hovy

[12] Indeed, when comparing against the less strict evaluation in Table 5, LSP$^{+A}$ incurs the smallest performance drop.

[13] Table 12 in the Appendix lists the consistency of all depths.

et al., 2009).

## 7 Conclusion

In this work, we bridge two powerful techniques in hypernym extraction: the pattern-based and prompt-based approach and use them as a diagnose tool to probe knowledge in BERT. We provide a thorough study of how patterns from the corpus-mining literature can be used to probe neural models. We find that LSP and DFP exhibit similar capacities, while anchored patterns bring consistent and significant benefits, suggesting a way to overcome challenging scenarios. In particular, we demonstrated clear benefits for rare hyponyms and abstract hypernyms, and an increase in the reliability of retrieved hypernyms under paraphrased prompts. This finding can direct future work on prompt design to extract robust and consistent hypernyms knowledge. The idea of anchoring prompts can be extended to other semantic relations such as part-of and synonyms to advance taxonomy induction and knowledge graph construction.

## 8 Limitations

**Effectiveness beyond noun-noun concepts**: we apply our method to hyponym-hypernym pairs over nouns in the general domain. This idea of anchored prompts can also be extended mine hypernyms for other parts-of-speech using patterns developed for text corpora (Chklovski and Pantel, 2004; Kozareva, 2014), as well as semantic relations beyond hyponyms-hyponyms, e.g., Part-Whole (Girju et al., 2003). We leave this exploration for future work.

**Time efficiency vs performance boost**: incorporating anchors boost the performance for hypernym extraction, however, we also need to consider that the performance improvements comes with additional time cost. Querying with anchored prompts require more computation when multiple anchors are used, although runtimes for the experiments in the paper are all very low.

**Hypernym diversity**: current work on extracting hypernyms with BERT predominantly considers single-word hypernyms and does not consider multi-word hypernyms or hypernyms that are not in the BERT vocabulary. Our work is no exception.

**Language diversity**: Most work in both hypernymy retrieval as well as language model prompting focuses on English, and as a consequence there is a lack of data sets in other languages. The extension of technologies to less well-resourced languages is a pressing direction for future research.

**Scale of Language Models** We focus on comparing different pattern structures with a single model, BERT-large. The behaviours of patterns under larger language models such as GPT3 (Brown et al., 2020) remains to be examined (Wei et al., 2022).

## References

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, Avignon, France. 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*, page 1–10. Association for Computational Linguistics.

Gabriel Bernier-Colborne and Caroline Barrière. 2018. CRIM at SemEval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 725–731, New Orleans, Louisiana. Association for Computational Linguistics.

Sai Abishek Bhaskar, Maximilian Köper, Sabine Schulte Im Walde, and Diego Frassinelli. 2017. Exploring multi-modal Text+Image models to distinguish between abstract and concrete nouns. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

Claudia Borg, Mike Rosner, and Gordon Pace. 2009. Evolutionary algorithms for definition extraction. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 26–32, Borovets, Bulgaria. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 33–40, Barcelona, Spain. Association for Computational Linguistics.

Barry Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior Research Methods*, 46:1119 – 1127.

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.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

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.

Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2005. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134.

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte m Walde. 2017. Contextual characteristics of concrete and abstract words. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Manuel Gimenes and Boris New. 2016. Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, 48:963–972.

Roxana Girju, Adriana Badulescu, and Dan Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 80–87.

Rebecca Green, Carol A. Bean, and Sung Hyon Myaeng. 2002. *The Semantics of Relationships: An Interdisciplinary Perspective*. Kluwer Academic Publishers, USA.

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*, pages 275–282, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*.

Eduard Hovy, Zornitsa Kozareva, and Ellen Riloff. 2009. Toward completeness in concept extraction and classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 948–957, Singapore. Association for Computational Linguistics.

Devansh Jain and Luis Espinosa Anke. 2022. Distilling hypernymy relations from language models: On the effectiveness of zero-shot taxonomy induction. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 151–156, Seattle, Washington. Association for Computational Linguistics.

Zornitsa Kozareva. 2014. Simple, fast and accurate taxonomy learning. In *Text Mining: From Ontology Learning to Automated Text Processing Applications*, pages 41–62, Cham. Springer International Publishing.

J. Lyons. 1977. *Semantics: Volume 2*. ACLS Humanities E-Book. Cambridge University Press.

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

Smaranda Muresan and Judith Klavans. 2002. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Daniela Naumann, Diego Frassinelli, and Sabine Schulte im Walde. 2018. Quantitative semantic variation in the contexts of concrete and abstract words. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 76–85, New Orleans, Louisiana. Association for Computational Linguistics.

Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A graph-based algorithm for inducing lexical taxonomies from scratch. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, page 1872–1877. AAAI Press.

Roberto Navigli, Paola Velardi, and Juana Maria Ruiz-Martínez. 2010. An annotated dataset for extracting definitions and hypernyms from the web. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia. Association for Computational Linguistics.

Marius Pasca. 2004. Acquisition of categorized named entities for web search. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management - CIKM '04*, page 137, Washington, D.C., USA. ACM Press.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2022. Discovering financial hypernyms by prompting masked language models. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 10–16, Marseille, France. European Language Resources Association.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

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.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

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*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Stephen Roller, Douwe Kiela, and Maximilian Nickel. 2018. Hearst patterns revisited: Automatic hypernym detection from large text corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–363, Melbourne, Australia. Association for Computational Linguistics.

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.

Timo Schick and Hinrich Schütze. 2019. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *AAAI Conference on Artificial Intelligence*.

Timo Schick and Hinrich Schütze. 2020. Rare words: A major problem for contextualized representation and how to fix it by attentive mimicking. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 65–75, Valencia, Spain. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

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.

Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada. 2021. Low-resource taxonomy enrichment with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2747–2758, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 1390–1397. AAAI Press.

# A Hypernym Evaluation

**Dataset Statistics** Table 9 presents the statistics on all datasets we used for experiments. We exclude hypernyms that are not included as single tokens in BERT vocabulary. The ratio of discarded $(x, y)$ pairs is lower than $1\%$ for most datasets, except for BLESS (30% is discarded) and CLSB (17% is discarded).

**Comparison with oracle anchors** To estimate the upper bound of anchored prompts, we treat siblings from WordNet (Miller, 1995) as oracle anchors and evaluate their effects on hypernym extraction. We select top five siblings with the highest rank of their path similarities calculated from WordNet, i.e., $\frac{1}{p(x,z)+1}$, where $p$ is the length of the shortest path between the $x$ and $z$ among their top two synsets. We use random sampling among siblings with the same score to select up to five anchors. The experimental results are presented in Table 10. We observe that using WordNet anchors can indeed lead to significant improvements in performance on datasets directly built from WordNet. For example, we observed large improvements for DIAG and LEDS when using WordNet anchors in combination with DFPA patterns. However, for other datasets, BERT anchors produce similar results as WorNet anchors. This highlights that with the improvement of anchor quality, anchoring prompts can unlock more hidden knowledge within BERT.

**Computational Resources** All experiments are conducted on single NVIDIA V100 GPU. A single run on each data set takes less than 2 hours, except for the large-scale dataset SHWARTZ, which takes nearly 24 hours on anchored prompts.

# B Consistency

## B.1 Pairwise consistency on close vocab

To compare our work with Ravichander et al. (2020) on pairwise probes using close vocab (nine hypernyms), we conduct the same experiments on DIAG dataset. Table 11 presents the results. The conclusion aligns with the open vocab set up: anchored patterns improve the consistency largely.

## B.2 Group consistency over different depths of hypernyms

Table 12 reports the group consistency across different depths of hypernyms.

| Dataset | #Hypon | #Hyper | #Pairs | WordNet Coverage (%) | Concreteness |
|---|---|---|---|---|---|
| BLESS (Baroni and Lenci, 2011) | 200 | 85 | 935 | 99.8 | 100 / 91.4 |
| DIAG (Ravichander et al., 2020) | 576 | 9 | 576 | 100 | 97.9/ 100 |
| CLSB (Devereux et al., 2014) | 508 | 232 | 1079 | 98.1 | 100/ 98.2 |
| SHWARTZ (Shwartz et al., 2017) | 11061 | 1101 | 12724 | 44.1 | 66.4/ 92.3 |
| LEDS (Baroni et al., 2012) | 1073 | 364 | 1262 | 100 | 83.7/ 79.2 |
| EVAL (Santus et al., 2015) | 621 | 348 | 953 | 99.8 | 88.1/ 83.4 |

Table 9: The statistics of datasets. WordNet Coverage is the coverage of hyponym-hypernym that are connected in WordNet on hypernyms hierarchy. Concreteness is the percentage of concrete hyponyms/hypernyms, measured by the concreteness rating from Brysbaert et al. (2014) for the shared vocab.

| | BLESS | | DIAG | | CLSB | | SHWARTZ | | EVAL | | LEDS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MRR | P@10 | MRR | P@10 | MRR | P@10 | MRR | P@10 | MRR | P@10 | MRR | P@10 |
| DFP | 23.6 | 42.4 | 42.6 | 66.8 | 39.8 | 67.5 | 6.3 | 12.8 | **24.0** | **46.7** | 32.6 | 60.1 |
| DFP$^{+A}$ | 25.7$^{+}$ | 47.2$^{+}$ | **45.5**$^{+}$ | **67.2**$^{+}$ | **42.3**$^{+}$ | **70.5**$^{+}$ | 5.9$^{+}$ | 13.6$^{+}$ | 22.1$^{+}$ | 43.3$^{+}$ | **35.7**$^{+}$ | **64.3**$^{+}$ |
| DFP$^{+A}_{oracle}$ ♯ | 23.9 | 41.9 | 65.4 | 84.6 | 41.2 | 68.2 | 8.9 | 15.6 | 23.3 | 45.1 | 37.7 | 66.2 |
| LSP | **27.1**$^{*}$ | **53.9**$^{*}$ | **45.5**$^{*}$ | 66.1 | 40.8 | 68.2 | 6.4 | **15.2**$^{*}$ | 17.3$^{*}$ | 39.5$^{*}$ | 33.4 | 60.5 |
| LSP$^{+A}$ | 26.5 | 53.2 | 42.8$^{+}$ | 62.7$^{+}$ | 40.4 | 67.7 | **6.5** | 14.9 | 17.0 | 38.1 | 34.0 | 61.6 |
| LSP$^{+A}_{oracle}$ ♯ | 26.2 | 49.6 | 65.6 | 85.8 | 41.9 | 68.3 | 9.1 | 18.8 | 18.7 | 40.5 | 37.1 | 66.3 |

Table 10: Main results on six hypernym extraction datasets with oracle anchors from WordNet. Bold number indicates the highest score per data set and metric. $^{*}$ indicates significant difference of LSP vs. DFP; $^{+}$ indicates significant difference wrt. the non-anchored counterpart (i.e., LSP vs LSP$^{+A}$ and DFP vs. DFP$^{+A}$). The ♯ symbol denotes that we report the average over 3 runs on sampled anchors from WordNet.

| Model | Patterns | | Accuracy | | |
|---|---|---|---|---|---|
| | Singular | Plural | Singular | Plural | Singular&Plural |
| Majority | - | - | 22.9 | 22.9 | 22.9 |
| BERT (Ravichander et al., 2020) [14] | A(n) X is a(n) Y | X are Y | 67.5 | 44.1 | 36.6 |
| DFP | A(n) X is a(n) Y. | X are Y. | 70.8 | 52.3 | 43.6 |
| DFP$^{+A}$ | A(n) X or Z is a(n) Y. | X or Z are Y. | 73.8 | 61.6 | **57.1** |
| LSP | Y such as a(n) X. | Y such as X. | 47.6 | 64.6 | 42.7 |
| LSP$^{+A}$ | Y such as a(n) X or Z. | Y such as X or Z. | 59.2 | 73.3 | 55.6 |

Table 11: Experimental results on pairwise singular-plural probes. X in singular patterns are singular format (e.g., car), while X in plural patterns are plural format (e.g., cars).

| Depth | #Instances | LSP | LSP$^{+A}$ | $\Delta$ | Hypernym Examples |
|---|---|---|---|---|---|
| 1 | 5 | 20.0 | 20.0 | 0.0 | transaction, conflict |
| 2 | 104 | 1.0 | 7.7 | 6.7 | object, group, relation, proceeding, battle |
| 3 | 352 | 2.8 | 3.7 | 0.9 | person, language, event, collection, trait |
| 4 | 1335 | 8.5 | 12.2 | 3.7 | band, organization, island, food, lake |
| 5 | 1572 | 6.7 | 11.1 | 4.4 | place, river, mountain, organisation, settlement |
| 6 | 4829 | 8.4 | 11.5 | 3.1 | film, village, company, animal, work |
| 7 | 1017 | 29.7 | 41.5 | 11.8 | vehicle, tool, country, plant, sport |
| 8 | 1773 | 9.4 | 14.8 | 5.4 | city, town, fruit, weapon, illness |
| 9 | 1110 | 32.0 | 36.2 | 4.2 | book, bird, magazine, mammal, tree |
| 10 | 348 | 28.2 | 32.5 | 4.3 | fish, ship, flower, airline, word |
| 11 | 15 | 6.7 | 6.7 | 0.0 | airplane, hawk, plane, vulture, murder |
| 12 | 12 | 16.7 | 25.0 | 8.3 | cancer, lizard, falcon, pine |
| 13 | 55 | 5.5 | 9.1 | 3.6 | human, pest, cat |
| 14 | 54 | 7.4 | 7.4 | 0.0 | horse |
| 16 | 2 | 50.0 | 100.0 | 50.0 | cattle |
| 17 | 2 | 0.0 | 0.0 | 0.0 | cow |

Table 12: Analysis on depth of hypernyms in WordNet. Column LSP and LSP$^{+A}$ are the group consistency (as in § 6.4) across depth. $\Delta$ is the gains from anchors (i.e., LSP$^{+A}$- LSP).