ALL Dolphins Are Intelligent and SOME Are Friendly: Probing BERT for Nouns' Semantic Properties and their Prototypicality

Marianna Apidianaki
Department of Digital Humanities
University of Helsinki
Helsinki, Finland
marianna.apidianaki@helsinki.fi

Aina Garí Soler
Université Paris-Saclay
CNRS, LISN
91400, Orsay, France
aina.gari@lisn.fr

Abstract

Large scale language models encode rich commonsense knowledge acquired through exposure to massive data during pre-training, but their understanding of entities and their semantic properties is unclear. We probe BERT (Devlin et al., 2019) for the properties of English nouns as expressed by adjectives that do not restrict the reference scope of the noun they modify (as in *red car*), but instead emphasise some inherent aspect (*red strawberry*). We base our study on psycholinguistics datasets that capture the association strength between nouns and their semantic features. We probe BERT using cloze tasks and in a classification setting, and show that the model has marginal knowledge of these features and their prevalence as expressed in these datasets. We discuss factors that make evaluation challenging and impede drawing general conclusions about the models' knowledge of noun properties. Finally, we show that when tested in a fine-tuning setting addressing entailment, BERT successfully leverages the information needed for reasoning about the meaning of adjective-noun constructions outperforming previous methods.

1 Introduction

Adjectival modification is one of the main types of composition in natural language (Baroni and Zamparelli, 2010; Guevara, 2010). Adjectives in attributive position[1] usually have a restrictive role on the reference scope of the noun they modify, limiting the set of things it refers to (e.g., *white rabbits* $\sqsubset$ *rabbits*). This property of adjectives has interesting entailment implications, generally leading to adjective-noun (AN) constructions where the entailment relationship with the head noun holds (AN $\models$ N) (Baroni et al., 2012). Entailment is directional (*white rabbit* $\models$ *rabbit* but *rabbit* $\not\models$ *white rabbit*) (Kotlerman et al., 2010), unless modification is not restrictive. When A is prototypical of the N it modifies (as in *soft silk, red lobster, small blueberry*), its insertion does not reduce the scope of N or add new information, but rather emphasises some inherent property (Pavlick and Callison-Burch, 2016). In these cases, N and AN denote the same set; they are in an equivalence relation (*red lobster* = *lobster*) and entailment is symmetric.

Masking Properties

SINGULAR	a balloon is [MASK].
PLURAL	balloons are [MASK].
SINGULAR + `usually`	a balloon is usually [MASK].
PLURAL + `usually`	balloons are usually [MASK].
SINGULAR+`generally`	a balloon is generally [MASK].
PLURAL + `generally`	balloons are generally [MASK].
SINGULAR + `can be`	a balloon can be [MASK].
PLURAL + `can be`	balloons can be [MASK].
`most` + PLURAL	most balloons are [MASK].
`all` + PLURAL	all balloons are [MASK].
`some` + PLURAL	some balloons are [MASK].

Masking Quantifiers

[MASK] balloons are colourful.	(ALL-MOST-SOME)
[MASK] balloons are large.	(SOME-SOME-FEW)
[MASK] balloons are round.	(MOST-SOME-NO)

Table 1: Cloze statements for the noun *balloon* with its properties (McRae et al., 2005) and quantifiers masked. Parentheses in the lower part of the table contain the gold quantifiers in (Herbelot and Vecchi, 2015).

The notion of prototypicality has a prominent place in the computational linguistics literature, mainly by reference to relationships between nouns (Roller and Erk, 2016; Vulić et al., 2017). The prototypicality of adjectives has been understudied and is absent from lexico-semantic resources such as WordNet (Fellbaum, 1998) and HyperLex (Vulić et al., 2017). Alongside the theoretical interest of this linguistic property and its impact on the entailment properties of AN constructions, identifying prototypical adjectives has interesting practical implications. It can serve to retrieve information

[1] Adjectives that appear immediately before the noun they modify and form part of the noun phrase (e.g., *white rabbit*), as opposed to adjectives in predicative position that occur after the noun (e.g., this *rabbit* is *white*).

Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 79–94

about the general concept (*silk, blueberry*) when queries include such AN pairs (*soft silk, small blueberry*), or to discard adjectives that do not add new information about the noun they modify in summarisation or sentence compression.

We investigate the knowledge that the BERT model (Devlin et al., 2019) encodes about nouns' inherent properties as described in AN constructions. Although pre-trained language models have been shown to encode rich factual and commonsense knowledge (Petroni et al., 2019; Bouraoui et al., 2020), little is known about their understanding of the properties of the involved entities. We specifically explore whether BERT encodes information about the prototypical properties of the class of objects denoted by a noun (for example, that *all lobsters are red* and *blueberries are small*). We use a set of collected norms that describe important concept features (McRae et al., 2005) and their associated quantifiers (Herbelot and Vecchi, 2015). We rely on these data to derive cloze statements that we use to query BERT about noun properties, and to train BERT-based classifiers predicting these properties. We furthermore fine-tune BERT for entailment and test it in a task that involves AN constructions (Pavlick and Callison-Burch, 2016). Our cloze task results show that BERT has only marginal knowledge of noun properties and their prevalence, but can still successfully detect cases where the addition of an adjective does not alter the meaning of a sentence and where entailment is preserved.

2 Related Work

Compositionality in AN constructions has been a central topic in distributional and formal semantics. Mitchell and Lapata (2010) derive the meaning representation of a composite phrase from that of its constituents by performing algebraic operations (addition and multiplication) on distributional word semantic vectors, while Baroni and Zamparelli (2010) and Guevara (2010) derive composite vectors through composition functions learned from corpus-harvested phrase vectors. In our work, we represent AN phrases by combining the contextualised BERT representations of A and N in sentences where they occur, using algebraic operations. We also investigate the extent to which the representations of A and N in an AN phrase capture its meaning, since token-level BERT embeddings encode information from the surrounding context.

We furthermore address the entailment relationship between N and ANs (N $\models$ AN). In the opposite direction (AN $\models$ N) entailment generally holds, i.e. almost all ANs entail their head noun (*red car* $\models$ *car*) (Baroni et al., 2012; Kober et al., 2021). Determining whether N $\models$ AN holds, however, depends on the semantic properties described by the adjective. We base this analysis on the AddOne dataset proposed by Pavlick and Callison-Burch (2016). AddOne consists of sentence pairs that contain AN phrases annotated for entailment through crowdsourcing. This simplified entailment task differs from the classical RTE task (Dagan et al., 2005) in that the premise and hypothesis differ by only one atomic edit (insertion of A). We use this task as a proxy for prototypicality based on the assumption that adjectives describing typical properties of a noun do not modify its scope (e.g., *lobster* $\models$ *red lobster*, but *car* $\not\models$ *red car*). Prototypicality has been addressed in the literature mainly with respect to nouns, i.e. the typical hyponyms in a specific semantic class (e.g., *dog* $\models$ *animal*), or member concepts that are most central to a category (Roller and Erk, 2016). Vulić et al. (2017) also address verb prototypicality in terms of how typical of an action a verb is (e.g., "Is TO RUN a type of TO MOVE?"). Our work extends this notion to adjectives describing noun properties in AN phrases.

On the probing side, previous work explores the factual and commonsense knowledge present in pretrained language models (LMs). The LAMA (LAnguage Model Analysis) probe proposed by Petroni et al. (2019) contains sets of facts from various knowledge sources.[2] Each fact is converted into a "fill-in-the-blank" cloze statement that is used to query the LM for a missing token. A model is considered to know a fact ([SUBJECT, `relation`, OBJECT] triple) if it can successfully predict masked tokens in cloze statements expressing this fact (e.g., DANTE `was born in` ____). The `HasProperty` relation in LAMA (extracted from ConceptNet) is similar to our relation of interest as it links nouns to adjectives describing their properties. ConceptNet contains 3,894 such pairs, but a close inspection of the data reveals several problematic cases (e.g., *informal both*, *divine forgive*, *ten 10*). Additionally, the cloze statements proposed for this dataset were automati-

[2]Relations between entities stored in Wikidata, common sense relations between concepts from ConceptNet (Speer and Havasi, 2012), and knowledge aimed for answering natural language questions in SQuAD (Rajpurkar et al., 2016).

cally extracted from Open Mind Common Sense (OMCS)[3] sentences and are often too long, including irrelevant information that might confuse the model.[4] Jiang et al. (2020) demonstrate the impact of prompt quality on LM probing but focus on relations involving encyclopedic knowledge (e.g., born/died in, profession, subclass). Bouraoui et al. (2020) also explore the knowledge BERT has about lexical, morphological and commonsense relations (e.g., hypernymy, meronymy, plural, cause-effect) through fine-tuning, but neither they address noun properties.

3 Datasets

McRae et al. (2005) dataset (MRD): Semantic feature norms are used in the field of psycholinguistics for studying human semantic representation and computation. MRD contains feature norms for 541 living and nonliving concepts collected from 725 participants in an annotation task. The annotators proposed features they thought were important for each concept, covering physical (perceptual), functional and other properties. Among the collected 7,258 concept-feature pairs, we find that a dolphin is *intelligent, friendly*, and *lives in oceans*; and that a chandelier is *hanging from ceilings* and *is made of crystal*. The number of annotators who proposed each feature is also provided. The dataset has been extensively used to investigate and improve the knowledge about object properties encoded by distributional models (Rubinstein et al., 2015), word embeddings (Lucy and Gauthier, 2017; Yang et al., 2018) and, more recently, contextual LMs (Forbes et al., 2019; Hasegawa et al., 2020). These studies do not focus on adjectival attributes but rather consider all proposed properties, or specific subsets such as visual properties. In our experiments, we explore noun properties through the "IS_ADJ" features of noun concepts present in MRD.

Herbelot and Vecchi (2015) dataset (HVD) : HVD adds an extra level of quantification annotations to the MRD norms. Three native speakers of English selected a natural language quantifier among [NO, FEW, SOME, MOST, ALL][5] for each concept-feature (*C,f*) pair, expressing the ratio of *C* instances having feature *f* (e.g., ALL guitars are musical instruments, but SOME guitars are electric). Quantification is important for semantic inference; it can serve to understand set relations (such as synonymy and hyponymy) and to derive logically entailed sentences. We use the HVD dataset to probe BERT for the prevalence of noun properties.

Pavlick and Callison-Burch (2016) dataset (AddOne) : The Addone dataset is focused on AN composition. It contains 5,560 sentence pairs involving an AN pair (s_N, s_{AN}) which have been manually annotated for entailment ($s_N \models s_{AN}$) by crowd workers. Addone sentence pairs differ by one atomic edit, the insertion of A (s_N: "There are questions as to whether our culture has changed.", s_{AN}: "There are questions as to whether our traditional culture has changed."). Sentences were collected from corpora of different genres and each pair was annotated with a score in a 5-point scale from 1 (contradiction) to 5 (entailment). Only the pairs with high agreement (same score assigned by 2 out of 3 annotators) were retained. We use the AddOne dataset to assess BERT's ability to detect entailment in AN constructions. The dataset comes with a pre-defined split into training, development and test sets (83/10/7%) which we use in our experiments addressing entailment.

4 Cloze Task Experiments

We use the `bert-base-uncased` and `bert-large-uncased` models pre-trained on the BookCorpus (Zhu et al., 2015) and on English Wikipedia (Devlin et al., 2019). The models are trained using a cloze task where tokens of the input sequence are masked and the models learn to fill the slots, and a binary classification objective where they need to predict whether a particular sentence follows a given sequence of words.

4.1 Cloze Task Probing for Properties

We retrieve adjective modifiers of nouns in MRD found in the IS_ADJ features describing a concept (*bouquet*: IS_*colourful*; *panther*: IS_*black*). There are 509 noun concepts with at least one IS_ADJ feature in MRD. We exclude features involving multi-word attributes (*coconut*: IS_*white_inside*, *raft*: IS_*tied_together_with_rope*) which we do not expect BERT to be able to predict.[6] The average

[3] https://github.com/commonsense/omcs

[4] For example: "To understand the event "The monkey ate some bananas.", it is important to know that Banana is [MASK]". The ground truth adjective in this case is *yellow*.

[5] NO and FEW labels were rarely used by the annotators and we consider them as describing cases of non typical attributes.

[6] We use "multi-word" to refer to attributes that involve multiple words separated with underscores. These are not necessarily idiomatic expressions.

# features	1	2	3	4	5	6	7	8	9
# nouns	98	124	97	76	60	35	12	6	1

Table 2: Number of nouns with a specific number of IS_ADJ features in MRD.

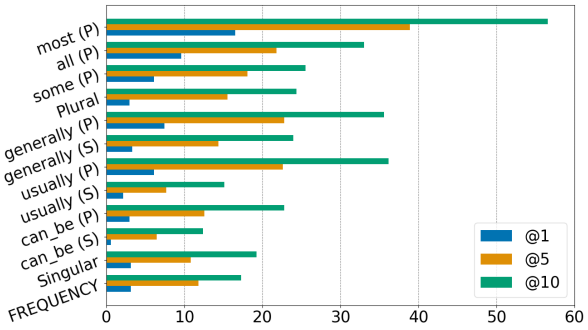

Figure 1: Accuracy at top-k with different query templates and BERT-`large`. The (S) and (P) markers denote singular and plural templates. Comparison to a frequency baseline.

number of features per noun is 3.12 (1,592 in total). Table 2 shows the number of nouns having a specific number of features. We define a set of templates and generate cloze statements for each noun that serve as our queries to probe BERT for these attributes. We define templates using both the singular and plural forms of a noun, as shown in Table 1.[7] We always use plural templates for 26 nouns given in plural form in MRD,[8] and singular templates for mass and uncountable nouns.[9] We evaluate the quality of the predictions made by BERT for each slot by checking the presence of ground-truth (reference) MRD adjectives in the top-1, top-5 and top-10 predictions ranked by probability. We compare BERT to a baseline that ranks by frequency all bigrams where a specific noun appears in the second position ("___ *bouquet*") in Google Ngrams (Brants and Franz, 2006), excluding bigrams that contain stop words and punctuation.

Similar to Ettinger (2020), we define accuracy as the percentage of items (nouns) for which the expected completion (one of the reference properties) is among the model's top-k predictions. The plot in Figure 1 shows the accuracy of predictions at top-k for BERT-`large`, with singular (S) and plural (P) queries. For $k = 1$, accuracy is very low, and naturally increases with $k = 5$ and $k = 10$. We observe that results differ considerably when different cloze statements are used for probing. Overall, BERT makes less accurate predictions with singular templates (e.g., a balloon usually is/can be [MASK]) than with plural templates (e.g., all/most balloons are [MASK]). Our assumption is that plural queries work better because they are more natural than singular ones,[10] and query naturalness seems to greatly influence prediction quality (Ettinger, 2020). The frequency baseline proposes correct properties for more nouns than the SINGULAR + `usually` and SINGULAR + `can be` templates. BERT-`large` gives slightly better results than BERT-`base` on this task. Best results are obtained with `most` + PLURAL queries (e.g., most balloons are [MASK]) where BERT-`large` proposes a correct attribute for 287 out of 509 nouns (56.4%). For BERT-`base`, best results are retrieved with PLURAL + `usually` queries (e.g., balloons are usually [MASK]), where a correct attribute is found in top-10 for 222 nouns. All results for BERT-`base` are given in Appendix B.1.

The results of this probing experiment suggest that BERT has marginal knowledge of noun properties as reflected in the MRD association norms, and highlight the difficulty to retrieve this kind of information from the representations using cloze task probes. This information (e.g., "bananas are yellow") is rarely explicitly stated in texts, in contrast to other types of lexical and encyclopedic knowledge (e.g., hypernymy: "a banana is a fruit") available in the Wikipedia texts that were used for model pre-training. Another issue with a cloze task evaluation for semantic properties is that contrary to structural properties (e.g., syntactic dependencies), there might be multiple correct answers for a query which might be partially covered by the resource used for evaluation. The quality of the resource, and the design of the task that served to gather the annotations, play an important role and should be taken into consideration for a fair interpretation of the results. We would, for example, expect annotators to propose a different set of properties if they were presented with cloze-type queries.

Although BERT does not always manage to predict the properties in MRD, we see from the quality of the proposed adjectives – quite high in some cases – that it encodes some knowledge about noun

[7] We use the plural form of nouns given by the `pattern` Python library and manually correct any errors.

[8] *beans*, *beets*, *curtains*, *earmuffs*, *jeans*, *leotards*, *mittens*, *onions*, *pajamas*, *peas*, *scissors*, *skis*, *slippers*, *shelves*, *sandals*, *bolts*, *gloves*, *nylons*, *boots*, *screws*, *pants*, *tongs*, *trousers*, *drapes*, *pliers*, *socks*.

[9] *rice*, *bread*, *football*.

[10] For example, "apples are red" has a higher frequency (919) than "an apple is red" (278) in Google Ngrams.

properties not present in the resource. For example, the predictions made for the probe "mittens are generally [MASK]" might not contain the gold MRD adjectives (*knitted*, *colourful*), but describe specific aspects such as their colour (*white, black, red, yellow*), shape and composition (*flat, thick, short, thin*), and the fact that they can be *removed* (i.e. a garment). Naturally, the quality of the predictions varies a lot across nouns. These often describe general knowledge about the described entity, as is the case in the predictions made for the query "all balloons are [MASK]": {*empty, free, flown, filled, lit, inflated, green, destroyed, closed, used*}. We also observe that most completions proposed by BERT are adjectives (more details in Appendix B.2).

BERT's predictions might contain synonyms of the adjectives present in MRD describing a correct property (for example, *deadly* instead of *lethal* for the noun *bomb*). We run an additional, more relaxed, evaluation where we also consider as correct the adjectives' synonyms in WordNet (Fellbaum, 1998).[11] As expected, we observe an increase in accuracy (cf. Appendix B.3). This highlights the limitation of the string matching approach and the need for a more flexible evaluation methodology.

We also conduct a human evaluation of the predictions made by BERT-`large` with the PLURAL + `most` template for 90 nouns. Two subjects (postgraduates in linguistics) annotated the top ten predictions made by the model as correct or wrong (independent of whether they were present in MRD or another resource). The task is different than the one that served to create MRD in that the annotators were not asked to propose adjectives for a noun, but instead they had to judge whether a prediction described some property of the noun. The micro-average inter-annotator agreement as measured with the Cohen's kappa (κ) coefficient was fair (0.39) (Landis and Koch, 1977; Artstein and Poesio, 2008). This demonstrates that deciding whether an adjective describes a property of a noun (instead of a state or some other marginal feature) is difficult for human annotators. We report the results of this evaluation in Appendix B.4.

Finally, we also investigate whether it is easier for BERT to propose correct properties for nouns that are not split into multiple tokens or WordPieces (Wu et al., 2016). The results (cf. Appendix B.5) confirm this intuition and show a slight decrease in the number of correctly predicted properties for nouns composed of multiple WordPieces.

[11] This is similar to synonym mapping in MT evaluation (Banerjee and Lavie, 2005; Marie and Apidianaki, 2015).

	Set A		Set B	
QUANT.	MRR	% of queries	MRR	% of queries
BERT-base				
ALL	0.203	79.57	0.207	75.25
MOST	0.167	64.47	0.138	55.07
SOME	0.188	79.06	0.156	70.67
BERT-large				
ALL	0.220	75.13	0.221	75.74
MOST	0.235	62.70	0.196	59.03
SOME	0.201	69.54	0.166	65.34

Table 3: MRR and proportion of queries in Sets A and B where a quantifier is predicted in top-10.

4.2 Cloze Task Probing for Quantifiers

In order to probe BERT for prototypicality, we split the AN pairs in HVD into two sets. Set (A) contains 788 pairs (for 386 nouns)[12] where the adjective describes a property that applies to most of the objects in the class denoted by the noun. These pairs are annotated with at least two ALL labels, or with a combination of ALL and MOST (*healthy banana* $\rightarrow$ [ALL-ALL-ALL]). We consider adjectives in Set (A) as describing prototypical properties of the nouns. Set (B), instead, contains 808 AN pairs (for 391 nouns) with adjectives describing properties of a smaller subset of the objects denoted by the noun. The labels assigned to these pairs contain SOME, FEW and NO's. For example, the annotations for *red apple* are [MOST-SOME-SOME], because there can be green and yellow apples.[13] We create cloze statements for the 788 pairs in Set (A), one for each pair, and for the 808 pairs in Set (B). A statement contains the noun in plural form and a masked slot for the quantifier (for example, "[MASK] *bananas are healthy*", "[MASK] *apples are red*"). We use each of the generated cloze statements to query BERT about prototypicality, checking whether it favours the expected (over the inappropriate) quantifier in its predictions. Given that BERT's predictions reflect the ranking of all words in the vocabulary according to whether they would be good fillers for the masked slot, we check the position of the quantifiers in the ranking.

[12] Some nouns have several AN pairs each.

[13] Note that a noun might be present in both Sets (A) and (B), depending on whether its ANs describe prototypical properties. We find, for example, *transparent jar* in Set (A), because all jars have this property, and *breakable jar* in Set (B), because not all jars can be easily broken.

We evaluate the predictions using Mean Reciprocal Rank (MRR) (cf. Appendix C). The results are shown in Table 3. The higher the MRR value is for a specific quantifier, the better its position in the ranking. If BERT encoded the knowledge needed to distinguish prototypical from other properties, we would expect ALL and MOST to be higher in the ranking produced for Set (A) queries, and SOME to come first in the ranking for Set (B) queries. We instead observe that ALL tends to occupy a higher position in the predictions for both sets. In Table 3, we also show the proportion of queries where a quantifier appears in the top-10 predictions. We observe that all three quantifiers are often proposed for queries in both sets. These results suggest that BERT does not distinguish properties on the basis of prototypicality.

When several quantifiers are found in the top-10 predictions, we also check their relative position in the ranking. We calculate the percentage of queries where BERT assigned a higher probability to the expected quantifiers, ranking them higher than the others. This corresponds to the "completion sensitivity test" proposed by Ettinger (2020), which serves to explore BERT's ability to prefer expected over inappropriate completions. No clear precedence pattern is detected: BERT-`base` assigns a higher probability to the expected (ALL) than to the incorrect completion (SOME) in 56% of Set (A) queries where both have been proposed; the inverse order is observed in 34% of queries in Set (B). We also run the sensitivity test on the ranking obtained for the whole vocabulary, but no meaningful patterns arise.[14] The use of a sensitivity threshold (Ettinger, 2020) turned out to be impractical in our setting because of the very low cloze probability assigned to the quantifiers in most cases. In our predictions, the definite article "*the*" is the most common top-1 prediction (in 85.7% of Set (A) queries, with an average probability of 0.629), followed by demonstrative and possessive determiners (e.g., *their*, *these*). The probability mass is concentrated on these words, hence the probability assigned to quantifiers is often close to zero.[15]

We also check whether the observed trends reflect the prior probability of the quantifiers and of the definite article in a large corpus. We approximate this using their frequency in Google Ngrams. We find that $freq(\text{THE}) > freq(\text{ALL}) > freq(\text{SOME}) > freq(\text{MOST})$. This is the same pattern obtained in our evaluation, with the exception of MRR results for BERT-`large`, where MOST is the highest ranked quantifier. This result suggests that BERT does not base prediction on the prevalence of noun properties but it, instead, largely follows the determiners' prior distribution.

5 Classification Experiments

We probe BERT representations for prototypicality in a classification setting, where models decide whether A describes a prototypical property of N. We use frozen embeddings (i.e. embeddings extracted from the pre-trained model) and fine-tuning.

5.1 Experimental Setup

Examples We consider as positive (prototypical) instances (`pos`) for this task AN phrases from HVD Set (A). As negative instances (`neg`) for a noun in Set (A), we use the AN pairs where it appears in Set (B). If $|\texttt{neg}| < |\texttt{pos}|$ for a noun, we collect additional negative instances from the ukWaC corpus (Baroni et al., 2009) where N is modified by an adjective A′ such that A′N $\notin$ HVD. We exclude cases where N is part of a compound (i.e. where it modifies another noun, as in *small sardine tin*).[16] We retain the most frequent ANs found for N in ukWaC as negative instances, until $|\texttt{neg}| = |\texttt{pos}|$.

Since common properties of nouns (e.g., *yellow banana*, *red strawberry*) are rarely explicitly stated in texts (Shwartz and Choi, 2020), we expect that the most frequent pairs found for a noun in ukWaC will not describe such properties. A manual exploration confirmed that the frequency-based filtering helps to retain good negative examples. The majority of the collected pairs do not describe prototypical properties (e.g., *useless pistol*, *organic celery*), with only a few (∼10) exceptions (e.g., *silvery minnow*). The final dataset contains 1,566 instances in total, 783 for each class (positive and negative).[17]

Representations For each AN in `pos` and `neg`, we obtain a BERT representation from a sentence (s_{AN}) in ukWaC where A modifies N. We pair s_{AN} with a sentence s_N where A has been automatically

[14] Appendix B.6 contains the quantifier precedence results.

[15] More details on determiners are given in Appendix B.6.

[16] We obtain the dependency parse of a sentence using `stanza` (Qi et al., 2020).

[17] We omitted five positive AN pairs because not enough negative instances were found for the noun in Set (B) or in ukWaC.

deleted (e.g., "Then shape into balls about the size of a small tangerine" vs. "Then shape into balls about the size of a tangerine"). We choose sentences where A is not modified by an adverb (e.g., *very small ant*, where removing *small* would result in an ungrammatical sentence). When no sentences are found for an AN (588 out of 1,566 cases), we use as s_{AN} the plural pattern from the cloze task experiments (e.g., *raspberries are edible*) and the plural noun alone as s_N (*raspberries*). When N is an uncountable noun, we use the singular pattern instead.[18] We also feed the bigram in isolation into BERT (configuration that we call ISO).

Data Split We keep aside 10% of the data as our development set and perform 5-fold cross-validation on the rest. To minimise the impact of lexical memorisation where the model learns that a word is representative of a specific class (Levy et al., 2015), we observe a full lexical split by adjective between the development set and the data used for cross-validation, and also between the training and the test set in each fold. As a result, adjectives found in the test set at each iteration have not been seen in the training or in the development set. This is done to avoid that the model memorises an adjective as describing a common or prototypical property of nouns; for example, *small* is a feature for 120 out of 509 nouns in MRD. The split allows to evaluate the capability of the model to generalise to unseen adjectives.

5.2 Embedding-based Classification

We expect the vector of an AN phrase involving a prototypical adjective (*red strawberry*) to be more similar to the vector of N (*strawberry*), than that of a phrase A′N involving an adjective that expresses a non typical property of N (*rotten strawberry*). We extract three contextualised embeddings from each layer of the BERT-`base` model that we use to compare the representation of an AN phrase to that of the head N:

1. an embedding for N in sentence s_N where N occurs without the adjective ($\overrightarrow{Ns_N}$);
2. an embedding for N in sentence s_{AN} which contains the adjective ($\overrightarrow{Ns_{AN}}$);
3. an embedding for A in s_{AN} ($\overrightarrow{As_{AN}}$).

We obtain an AN representation ($\overrightarrow{AN}$) by combining the vectors pairwise: $\overrightarrow{Ns_N}$ and $\overrightarrow{Ns_{AN}}$; $\overrightarrow{Ns_N}$ and $\overrightarrow{As_{AN}}$; $\overrightarrow{Ns_{AN}}$ and $\overrightarrow{As_{AN}}$, using different composition operations: average, concatenation, difference, multiplication, and addition. We also experiment with the token-level contextualised representations $\overrightarrow{As_{AN}}$ and $\overrightarrow{Ns_{AN}}$ only, which we expect to also encode information about the noun and the adjective in the AN, respectively, since they occur in the same context. We use the different AN representations as features for a logistic regression classifier. Additionally, we calculate the cosine similarity and euclidean distance between the representation of a noun ($\overrightarrow{Ns_N}$ or $\overrightarrow{Ns_{AN}}$) and $\overrightarrow{AN}$ obtained through the vector combinations and composition operations described above, and feed them to the classifier as individual features or in combination. For comparison, we also run experiments using static word2vec (Mikolov et al., 2013) and fastText (Mikolov et al., 2018) embeddings as features, creating $\overrightarrow{AN}$ with the word embeddings $\overrightarrow{N}$ and $\overrightarrow{A}$, and using $\overrightarrow{A}$ alone. For each type of representation (BERT, word2vec, fastText), we select the configuration with the highest average accuracy on the development set over the five cross-validation runs. In Table 4, we report the average scores obtained on the test sets of the five folds for these configurations. Precision, recall and F1-score show how good a model is at detecting AN pairs that involve a prototypical adjective. As baselines, we provide results for a model that always predicts prototypicality (ALL-PROTO), and a model that assigns the majority label found in the training set at each fold (MAJORITY).

Model	Acc	F1	P	R
BERT	**0.658**	0.648	**0.676**	0.633
BERT (ISO)	0.586	0.548	0.605	0.506
fastText	0.593	0.481	0.639	0.411
word2vec	0.559	0.455	0.601	0.372
ALL-PROTO	0.507	**0.672**	0.507	**1.000**
MAJORITY	0.473	0.524	0.390	0.800

Table 4: Average accuracy, F1-score, precision (P) and recall (R) of embedding-based classifiers on HVD in the cross-validation experiment across five folds.

In terms of accuracy, BERT obtains the best results on this task (0.658) when cosine similarity and euclidean distance between $\overrightarrow{Ns_N}$ and $\overrightarrow{Ns_N}$+$\overrightarrow{Ns_{AN}}$ at the last (12th) layer are used as features. The simple ALL-PROTO baseline obtains the highest F1 score (0.672) but gets low accuracy in this balanced dataset. Using sentences containing the AN is more effective than feeding the AN

[18] Using sentences created with these patterns for all ANs hurts performance compared to the setting where sentences gathered from corpora are used.

Model	Acc	F1	P	R
BERT-CLS	0.696	**0.654**	0.763	**0.582**
BERT-TOK	**0.697**	0.646	**0.778**	0.561
BERT-CLS (ISO)	0.604	0.503	0.654	0.424
BERT-TOK (ISO)	0.636	0.591	0.701	0.539

Table 5: Average accuracy, F1 score, precision and recall in the cross-validation experiment across five folds for a BERT model fine-tuned on HVD using the CLS and TOK approaches.

bigram in isolation into BERT (ISO). Static representations, especially word2vec, perform worse than BERT but still manage to beat the baselines in terms of accuracy. The best configuration for word2vec and fastText was the use of the adjective representations ($\overrightarrow{A}$) as features, which shows that the models do not manage to extract the information needed for assessing prototypicality from the different $\overrightarrow{N}$ and $\overrightarrow{A}$ combinations. Instead, the best strategy is to learn the tendency of an adjective to be prototypical. When evaluated on unseen adjectives in our test sets, they base prototypicality judgments on the similarity of these adjectives to the ones seen in the training set. We observe a high variation in accuracy and F1 scores across folds for all models. For BERT, F1 scores range from 0.553 to 0.740 and the range is even larger for the fastText-based model (from 0.310 to 0.747). This suggests that prototypicality is not easy to detect for all AN pairs. Overall, BERT embeddings seem to be a better fit for estimating prototypicality than static representations. We report the detailed results by layer, and the best configurations per $\overrightarrow{AN}$ and composition type, in Appendix D.

5.3 Fine-tuning BERT

We compare our results with frozen embeddings to the performance of BERT fine-tuned for the prototypicality task. Specifically, we feed into BERT the two sentences in each (s_N, s_{AN}) pair separated by the [SEP] token. We experiment with a classifier on top of the [CLS] token, as is typically done in sentence-pair classification tasks (we call this approach BERT-CLS); and with a classifier on top of the concatenation of two token representations: ($\overrightarrow{Ns_N}$, $\overrightarrow{As_{AN}}$), ($\overrightarrow{Ns_N}$, $\overrightarrow{Ns_{AN}}$), ($\overrightarrow{Ns_N}$, $\overrightarrow{As_{AN}} + \overrightarrow{Ns_{AN}}$) (our BERT-TOK approach). The two classification heads consist of a linear layer with softmax and are trained with a cross entropy loss. We fine-tune each model for 3 epochs with 0.1 dropout, and choose the learning rate based on the accuracy on the development set. Results of this experiment are found in Table 5. BERT-CLS and BERT-TOK ($\overrightarrow{Ns_N}$, $\overrightarrow{As_{AN}}$) perform comparably on this task and obtain better results than embedding-based models (Table 4), with 0.697 accuracy. As in the experiment described in Section 5.2, using sentences yields better results than only feeding the AN (ISO).

6 Entailment in AN Constructions

6.1 Task Description

AN constructions are often in a forward entailment relation with the head noun (*white rabbit* $\models$ *rabbit*) (Baroni et al., 2012). Whether backward entailment holds depends on the properties of N described by A. For example, a *car* is not always *red* (the label would be "Unknown"), while *lobster* always entails *red lobster*. We explore BERT's capability to identify the AN cases where backward (N $\models$ AN) entailment holds using the Addone dataset (Pavlick and Callison-Burch, 2016).

We fine-tune BERT on Addone to assess whether it captures the entailment relationship in AN constructions. BERT has shown high performance in other textual entailment tasks (Devlin et al., 2019), but the Addone dataset has proved challenging for other models relying on recurrent architectures. We follow Pavlick and Callison-Burch (2016) and use Addone for a binary classification task, with the labels ENTAILMENT (for forward entailment and equivalence) and NOT ENTAILMENT (encompassing the contradiction, independence and reverse entailment relations). Similarly to the fine-tuning approach described in Section 5.3, we feed into BERT the two sentences in each pair (s_N, s_{AN}) separated by the special [SEP] token. We again use the CLS and TOK classification heads. We fine-tune the model for 5 epochs with 0.1 dropout and select the learning rate based on the F1 score calculated over the actual ENTAILMENT cases on the development set.[19]

6.2 Results

Results of our experiments on Addone are presented in Table 6. We include results reported by Pavlick and Callison-Burch (2016) (P&CB) for comparison. The MAJ and MAJ-BY-ADJ baselines assign the majority class in the training set (NON-

[19] We use F1 score as a criterion, and not accuracy, because the Addone dataset is highly imbalanced (only 23% of the instances belong to the ENTAILMENT class).

Model	Acc	F1	P	R
Human (P&CB)	0.933	0.730	0.840	0.640
MAJ-BY-ADJ (P&CB)	**0.922**	0.680	**0.860**	0.560
MAJ (P&CB)	0.853	-	-	-
BERT-TOK	0.912	**0.696**	0.709	0.684
BERT-CLS	0.147	0.257	0.147	**1.000**
RNN (P&CB)	0.873	0.510	0.600	0.440

Table 6: Results on the Addone test set. Best results for each metric are highlighted in boldface.

ENTAILMENT) and the majority class proposed for each adjective in the training set, respectively. We also report the human performance on this task as an upper bound, and compare to the best-performing model in Pavlick and Callison-Burch (2016) which relies on a RNN architecture (Bowman et al., 2015). BERT-CLS fails to learn the information needed for the task and predicts the ENTAILMENT label for all instances. This explains the low scores obtained with this model, since the majority label in this dataset is NON-ENTAILMENT. The default fine-tuning strategy used for textual entailment with BERT is, thus, not suitable for addressing cases of compositional entailment in the Addone dataset. It is much more effective to use the representations of the specific words that determine sentence entailment: BERT-TOK ($\overrightarrow{Ns_N}$, $\overrightarrow{As_{AN}}$) obtains higher results than the previous best model (RNN) and beats the MAJ baseline in terms of accuracy, as well as MAJ-BY-ADJ in terms of F1 and recall. This shows that BERT leverages the AN relations that are needed to solve this NLI task better compared to RNNs.

7 Discussion

Retrieving prototypical knowledge about entities is a real challenge for distributional models, not necessarily because of the models themselves and how advanced they are, but because this information is rarely stated in texts. This is described in the literature as the "reporting bias" phenomenon (Gordon and Van Durme, 2013) which poses challenges on knowledge extraction. According to this phenomenon, rare actions or properties are over-represented in texts at the expense of trivial ones. Shwartz and Choi (2020) show that the generalisation capability of pre-trained language models allows them to better estimate the plausibility of frequent but unspoken actions, outcomes and properties than previous models, but that they also tend to overestimate that of the very rare, amplifying the bias that already exists in their training corpus. In this study, using methodology commonly used for probing contextual models, we have precisely explored whether retrieving knowledge about noun properties constitutes a challenge for these models, or whether they manage to retrieve this knowledge due to their impressive generalisation capabilities.

Since prototypical properties are often visual or perceptual, in future work we plan to combine text and visual features (Silberer et al., 2013; Lazaridou et al., 2015; Lu et al., 2019; Li et al., 2019, 2021) for retrieving noun properties, in order to see how BERT can predict these properties when it has access to images alongside text. Another goal is to collect evaluation data using cloze task queries in specifically designed crowdsourcing tasks.

8 Conclusion

We have conducted a thorough investigation of the information encoded by BERT about nouns' intrinsic properties. Using datasets specifically compiled for psycholinguistics studies, we probed BERT for noun properties and their prototypicality, as well as for the entailment relationship involving AN constructions which indicates possible generalisations to the entire class denoted by the noun. Our results show that information about noun properties, as described in word association norms, is hard to retrieve using cloze tasks. We discuss the limitations of semantic cloze tasks evaluation against existing resources, and the need for more flexible evaluation scenarios. However, knowledge about properties can still be leveraged by BERT in a classification setting where the model is exposed to examples specifically encoding this information.

We make our code and datasets available to promote further research in this direction.[20]

Acknowledgements

This work has been supported by the French National Research Agency under project ANR-16-CE33-0013. The work is also part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 771113). We thank the anonymous reviewers for their helpful feedback and valuable suggestions.

[20]The code and datasets are available at this URL: https://github.com/ainagari/prototypicality

References

Ron Artstein and Massimo Poesio. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. 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, Avignon, France. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193, Cambridge, MA. Association for Computational Linguistics.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing Relational Knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7456–7463, New York, NY, USA. AAAI Press.

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

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. In *LDC2006T13*, Philadelphia, Pennsylvania. Linguistic Data Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

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.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation*, pages 1753–1759, Montreal, Canada.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Emiliano Guevara. 2010. A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. In *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*, pages 33–37, Uppsala, Sweden. Association for Computational Linguistics.

Mika Hasegawa, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2020. Word Attribute Prediction Enhanced by Lexical Entailment Tasks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5846–5854, Marseille, France. European Language Resources Association.

Aurélie Herbelot and Eva Maria Vecchi. 2015. From concepts to models: some issues in quantifying feature norms. *Linguistic Issues in Language Technology (LiLT)*, 2(4).

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.

Thomas Kober, Julie Weeds, Lorenzo Bertolini, and David Weir. 2021. Data Augmentation for Hypernymy Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1034–1048, Online. Association for Computational Linguistics.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. 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.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint:1908.03557*.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, Online. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32, Vancouver, Canada.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

Benjamin Marie and Marianna Apidianaki. 2015. Alignment-based sense selection in METEOR and the RATATOUILLE recipe. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 385–391, Lisbon, Portugal. Association for Computational Linguistics.

Ken McRae, George Cree, Mark Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37:547–59.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint:1301.3781v3*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional Entailment in Adjective-Nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173, Berlin, Germany. 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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Stephen Roller and Katrin Erk. 2016. Relations such as Hypernymy: Identifying and Exploiting Hearst Patterns in Distributional Vectors for Lexical Entailment. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2163–2172, Austin, Texas. Association for Computational Linguistics.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How Well Do Distributional Mod-

els Capture Different Types of Semantic Knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 726–730, Beijing, China. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of Semantic Representation with Visual Attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 572–582, Sofia, Bulgaria. Association for Computational Linguistics.

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

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment. *Computational Linguistics*, 43(4):781–835.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint:1609.08144*.

Yiben Yang, Larry Birnbaum, Ji-Ping Wang, and Doug Downey. 2018. Extracting Commonsense Properties from Embeddings with Limited Human Guidance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 644–649, Melbourne, Australia. Association for Computational Linguistics.

A Masking Templates

Table 7 contains the templates that were used to construct the singular and plural queries for different nouns. SINGULAR_NOUN and PLURAL_NOUN are placeholders for the nouns in singular and plural form, respectively.

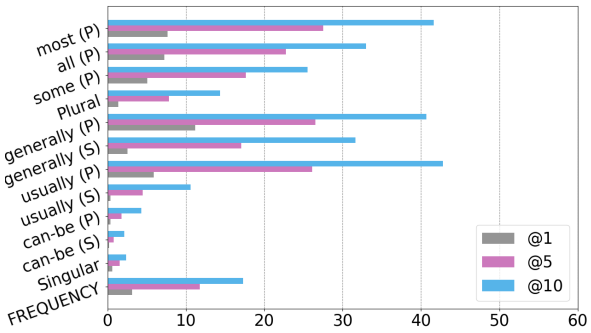

Figure 2: Accuracy at top-*k* for BERT-`base`. (S) and (P) stand for singular and plural templates. We compare to the results of a frequency baseline.

We generated the queries for the quantifiers using the template:

[MASK] `PLURAL-NOUN are ADJECTIVE .`

An example query generated using this template is:

[MASK] balloons are colourful .

B Additional Masking Results

In this section of the Appendix, we present in more detail the results that we obtained in the masking experiments for noun properties and quantifiers.

B.1 Property Masking Results

The plot in Figure 2 shows the accuracy of predictions at top-*k* for BERT-`base`. Figures 3 and 4 show the average recall at the top-1, top-5 and top-10 positions of the ranked BERT-`base` and `large` predictions, when using sentences constructed with the templates that correspond to the labels on the x axis. Average is calculated over the words for which at least one correct attribute is found at the specific rank.

B.2 Adjectives in BERT Predictions

Figure 5 shows the proportion of adjectives predicted by BERT-`large` using different templates. We count as adjectives all words that have a synset of this part of speech in WordNet. We observe that the majority of predictions pertain to this part-of-speech. Fewer adjectives are proposed for queries of the form SINGULAR + `can be` (e.g., a balloon can be [MASK]), where BERT tends to favour verbs in the past participle form.

B.3 WordNet-based Evaluation

Figure 6 presents the results of our more relaxed evaluation, which includes the WordNet synonyms of the adjectives in MRD. Specifically, we expand

SINGULAR	`a SINGULAR_NOUN is [MASK].`
PLURAL	`PLURAL_NOUN are [MASK].`
SINGULAR + `usually` PLURAL + `usually`	`a SINGULAR_NOUN is usually [MASK].` `PLURAL_NOUN are usually [MASK].`
SINGULAR + `generally` PLURAL + `generally`	`a SINGULAR_NOUN is generally [MASK].` `PLURAL_NOUN are generally [MASK].`
SINGULAR + `can be` PLURAL + `can be`	`a SINGULAR_NOUN can be [MASK].` `PLURAL_NOUN can be [MASK].`
PLURAL + `most`	`most PLURAL_NOUN are [MASK].`
PLURAL + `all`	`all PLURAL_NOUN are [MASK].`
PLURAL + `some`	`some PLURAL_NOUN are [MASK].`

Table 7: Masking templates with the noun in singular and plural form.

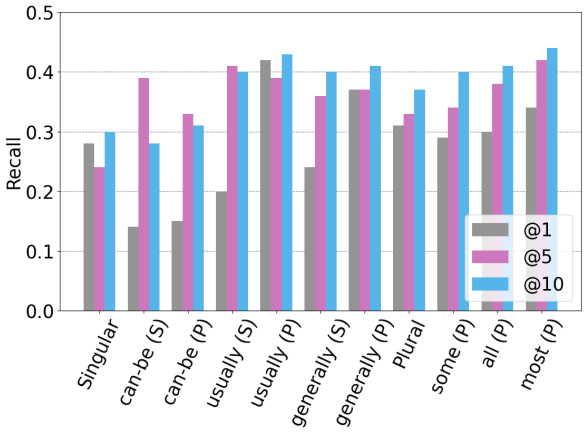

Figure 3: Average recall of MRD adjectives in the top-k predictions made by BERT–`base`.

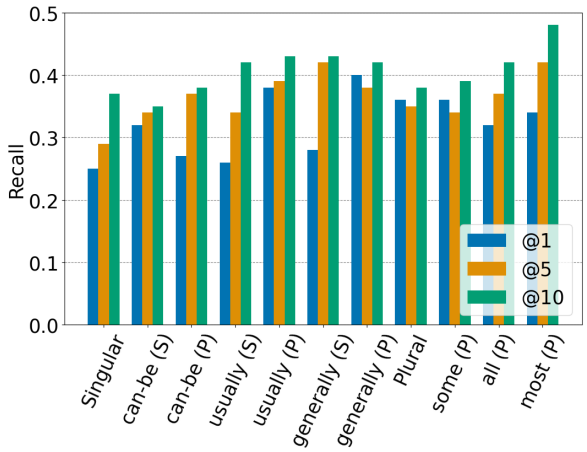

Figure 4: Average recall of MRD adjectives in the top-k predictions made by BERT–`large`.

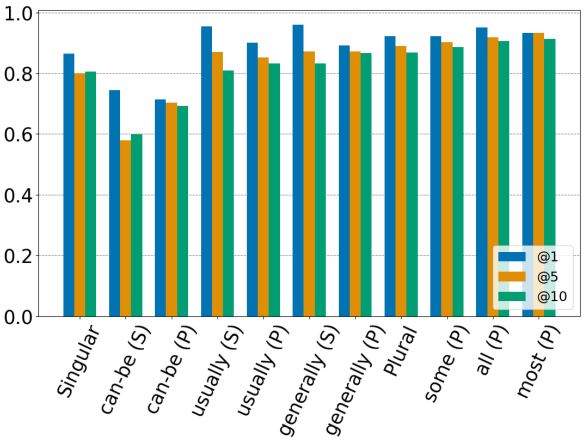

Figure 5: Proportion of adjectives among BERT-`large` predictions at different k.

the set of adjectives proposed in MRD for a noun (i.e. our reference) with their synonyms found in WordNet, and consider them all as correct. The lighter shades in the Figure show the improvement in accuracy at top-k with respect to our previous results (darker shades). This shows that the model sometimes predicts correct properties that cannot be captured in an evaluation based solely on string matching.

B.4 Manual Evaluation of Predicted Properties

In Table 8, we present the results of the manual evaluation. We show the number of properties that were marked as correct by each annotator (A#1 and A#2). We also report the number of properties for which the annotators agreed they were correct (Both). In the last column, we compare to the number of predictions that were found to be correct when evaluated solely against the reference properties in MRD. Since MRD has a different number of properties per noun – often fewer than ten – we indicate in parentheses the upper bound that could

	Manual evaluation			
	A#1	**A#2**	**Both**	**MRD**
@1	57	45	39	15 (90)
@5	228	166	130	50 (327)
@10	426	291	224	89 (355)

Table 8: Number of properties predicted by BERT-`large` with the "PLURAL + `most`" The last column shows the number of correct predictions when evaluating against properties in MRD. The upperbound for this evaluation is given in parentheses.

be reached if all reference properties for a noun were correctly predicted.

Agreement between the annotators is fair (0.39), which demonstrates that deciding whether an adjective describes a property of a noun is difficult. The annotators highlighted that there are some adjectives that BERT often proposes for nouns describing a specific class; for example, *nocturnal*, *solitary* and *shy* were proposed for different animals. We also find different colours proposed for *butterfly* (*white*, *black*, *brown*, *green*, *blue*) instead of the adjective *colourful* which describes a more general property of the insect (and which is one of its features in MRD).

B.5 Impact of Word Splitting on Performance

BERT uses WordPiece tokenization (Wu et al., 2016). The most frequent words are represented with a single token, but other words are split into multiple wordpieces. We investigate the impact of wordpiece splitting on the results. We run this analysis with BERT-`large` predictions at top-10. We classify nouns into two classes, `correct` and `incorrect`, depending on whether at least one of their properties was correctly predicted. We compare the proportion of multi-piece nouns (MPs) in the two classes using χ^2 tests. We observe a significant difference ($\alpha = 0.05$) with 4 out of our 11 templates. Table 9 contains the p-values and effect sizes (Cramer's V) for these templates. We also report the proportion of MPs in the `incorrect` class, which is slightly higher than that over all nouns in our dataset.[21] The effect size values are also weak, suggesting that word splitting has only a slight negative effect on BERT's performance on this cloze task.

[21]The number of MP nouns in singular and plural templates differs because plural forms of the nouns are more often split into multiple tokens.

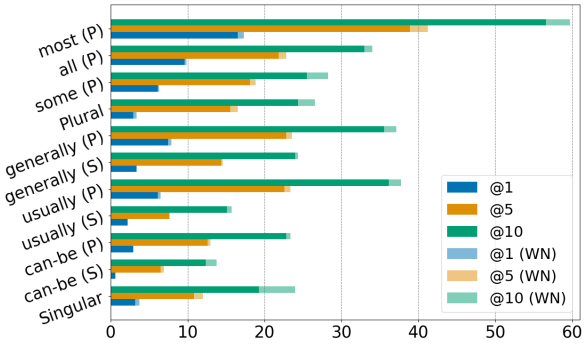

Figure 6: Improvement in accuracy at top-k for BERT-`large` predictions in the WordNet-based (WN) evaluation.

B.6 Additional Quantifier Probing Results

Table 10 shows the relative position of correct vs. incorrect quantifiers when they both appear in the top-10 predictions (columns 2 and 3), and in the ranking for the whole vocabulary (column 4). Correct (expected) completions for Set (A) queries are ALL and MOST; for Set (B), the correct answer is SOME.

The symbol ">" in the first column denotes precedence of a quantifier over another (i.e. higher probability). Column 3 shows the number of queries for which the two quantifiers were proposed in top-10, which served to calculate the proportion in column 2. ALL and SOME were, for example, proposed by BERT-`base` in top-10 for 532 Set (A) queries. ALL had a higher probability than SOME in 56% of these queries.

Table 11 shows the ranking results for other determiners (*the*, *these*, *their*) that are found in the first (top-1) position.

C Mean Reciprocal Ranking

Equation 1 contains the formula for Mean Reciprocal Ranking (MRR). RR measures the reciprocal of the rank of the first correct answer, and MRR is the average across queries. We use MRR in Section 4.2 to evaluate quantifier prediction. $|Q|$ corresponds to the number of queries.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (1)$$

D Detailed Embedding-based Classification results

In Table 13, we report the best results obtained on the HVD development set for each type of BERT-

	p-value	effect size (Cramer's V)	% of MPs with no correct predictions	total % MPs
SINGULAR + `usually`	0.011	0.11	30.7%	
SINGULAR + `generally`	0.038	0.09	30.7%	28.2%
SINGULAR + `can be`	0.002	0.14	30.6%	
PLURAL	0.03	0.03	59.2%	56.4%

Table 9: Statistics of the χ^2 tests that showed a significant association between (a) a noun being multi-piece (MP), and (b) BERT predicting wrong properties for the noun (as evaluated against MRD). The last column shows the proportion of MP nouns in singular and plural templates.

	Results @10		Results @|V|
BERT-base			
QUANTIFIERS	**% queries**	**# queries**	**% queries**
Set (A)			
ALL > SOME	56.02%	532	53.43%
MOST > SOME	49.89%	451	37.31%
Set (B)			
SOME > ALL	34.48%	467	40.1%
SOME > MOST	83.87%	31	65.47%
BERT-large			
QUANTIFIERS	**% queries**	**# queries**	**% queries**
Set (A)			
ALL > SOME	55.19%	462	55.08%
MOST > SOME	58.00%	431	44.16%
Set (B)			
SOME > ALL	33.41%	449	35.02%
SOME > MOST	48.00%	25	51.98%

Table 10: Relative position of correct vs. incorrect quantifiers when they both appear in the top-10 predictions made by each model, and in the ranking for the whole vocabulary.

	Set (A)		Set (B)	
BERT-base				
	% queries	Avg. Prob.	% queries	Avg. Prob.
the	85.7%	0.629	90.5%	0.662
these	6.2%	0.350	3.8%	0.375
their	0.8%	0.523	1.0%	0.686
BERT-large				
	% queries	Avg. Prob.	% queries	Avg. Prob.
the	74.3%	0.664	79.9%	0.716
these	4.2%	0.419	3.1%	0.510
their	0.4%	0.398	0.7%	0.572

Table 11: Proportion of queries in each set where the determiners *the*, *these* and *their* are found at the first position in the ranking. We also report the average probability assigned to a determiner when found in this position.

based $\overrightarrow{AN}$ representation and composition operation. The combination of Ns_N and Ns_{AN} clearly outperforms the other vector combinations. Using the adjective token-level representation alone

$\overrightarrow{AN}$ type	Acc	composition	Acc
Ns_N, Ns_{AN}	0.712	addition	0.712
As_{AN}	0.675	difference	0.667
Ns_N, As_{AN}	0.667	concatenation	0.660
Ns_{AN}, As_{AN}	0.665	average	0.650
Ns_{AN}	0.613	multiplication	0.611

Table 12: Highest average accuracy obtained by the different types of AN representation (left) and composition operations (right) with BERT embedding-based classifiers on the HVD development set.

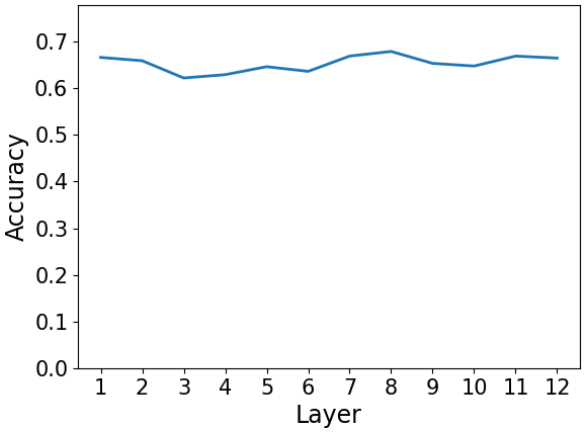

Figure 7: Highest average accuracy obtained by the embedding-based classifier on the HVD development set at every BERT layer.

($\overrightarrow{As_{AN}}$) also yields good results, definitely higher than $\overrightarrow{Ns_{AN}}$. In terms of composition functions, addition is the best performing operation for this task and multiplication the least useful.

Figure 7 shows the highest average accuracy obtained by each BERT layer on the HVD development set in these experiments.

$\overrightarrow{AN}$ **type**	**Composition**	**Layer**	**Similarity**	**Accuracy**
Ns_N, Ns_{AN}	addition	12	cosine & euclidean (Ns_N, $Ns_N + Ns_{AN}$)	0.712
As_N	-	8	-	0.675
Ns_N, As_{AN}	difference	12	-	0.667
Ns_{AN}, As_{AN}	difference	12	-	0.665
Ns_{AN}	-	11	cosine (Ns_N, Ns_{AN})	0.613
Composition type	$\overrightarrow{AN}$ **type**	**Layer**	**Similarity**	
addition	Ns_N, Ns_{AN}	12	cosine & euclidean (Ns_N, $Ns_N + Ns_{AN}$)	0.712
difference	Ns_N, As_{AN}	12	-	0.667
concatenation	Ns_{AN}, As_{AN}	7	-	0.660
average	Ns_N, As_{AN}	5	-	0.650
multiplication	Ns_N, As_{AN}	7	euclidean (Ns_N, $Ns_N \odot Ns_{AN}$)	0.611

Table 13: Best configurations for each type of $\overrightarrow{AN}$ (top) and composition operations (bottom) with BERT embedding-based classifiers. These are identified based on the accuracy obtained on the HVD development set.