Adjective Scale Probe: Can Language Models Encode Formal Semantics Information?

Wei Liu[1], **Ming Xiang**[2], **Nai Ding**[1]*

[1]College of Biomedical Engineering and Instrument Sciences, Zhejiang University
[2]Department of Linguistics, The University of Chicago
liuweizju@zju.edu.cn, mxiang@uchicago.edu, ding_nai@zju.edu.cn

Abstract

It is an open question what semantic representations transformer-based language models can encode and whether they have access to more abstract aspects of semantic meaning. Here, we propose a diagnostic dataset to investigate how well language models understand the degree semantics of adjectives. In the dataset, referred as the Adjective Scale Probe (ASP), we semi-automatically generate 8 tests of Natural Language Inference (NLI) questions to test 8 key capabilities of adjective interpretation. We apply the ASP dataset to evaluate the performance of 3 language models, i.e., BERT, DeBERTa, and T0. It is found that language models perform below the majority baseline for most tests of the ASP, even when the models have been fine-tuned to achieve high performance on the large-scale MNLI dataset. But after we fine-tune the pre-trained models on a subset of the ASP, DeBERTa can achieve high performance on the untrained adjectives and untrained tests, suggesting that DeBERTa may have captured degree semantic information of adjectives through pre-training but it needs specific training data to learn how to apply such information to the current tasks. In sum, the ASP provides an easy-to-use method to test fine-grained formal semantic properties of adjectives, and reveals language models' abilities to access formal semantic information.

Introduction

Transformer-based language models have approached or even surpassed human performance in many linguistic tasks (Devlin et al. 2019; He, Gao, and Chen 2021), but they are also known to be susceptible to adversarial attacks (Wallace et al. 2019; Lin, Zou, and Ding 2021) and perform poorly on some diagnostic datasets (Naik et al. 2018; McCoy, Pavlick, and Linzen 2019). Therefore, it remains debated to what extent the models truly understand language and what types of semantic information such models can encode. Language models represent word meaning as vectors calculated over the contexts that a word appears. This has the advantage to capture nuanced relations between words since similar words tend to occur in similar contexts (Yenicelik, Schmidt, and Kilcher 2020; Miaschi et al. 2020). It is unclear, however, whether such models can encode more abstract aspects of the word meaning (Bender and Koller 2020; Bisk et al. 2020). Here, we investigate whether language models can encode more abstract and complex semantic information by examining models' understanding of a major class of words, i.e., gradable adjectives[1].

On the one hand, gradable adjectives have highly context-sensitive meaning, since the applicability of an adjective to a noun varies from context to context. The utterance "*John is tall*" may be true when *John* is compared to other high school students, but the same utterance could be false when *John* is compared to a group of basketball players. On the other hand, in formal semantics, the degree semantics analysis (Cresswell 1976; Stechow 1984; Heim 2000; Kennedy and McNally 2005; Kennedy 2007) of adjectives postulates an abstract semantic core underlying the meaning of all adjectives in all contexts, which boils down to a few crucial parameters. Briefly, an utterance in the form "*X is Adj*" would map the object *X*, via a measure function, to a degree on a relevant scale defined by a particular dimension associated with the adjective. And the utterance is true if and only if the degree of *X* on the scale is larger than a contextually salient threshold. For example, for the sentence "*The elephant is heavy*", a measure function takes the argument *the elephant* and maps it to its weight (i.e. a degree d_{weight}) on a weight scale, and it says that the weight of *the elephant* is greater than a contextually determined threshold on the same scale (e.g. the average weight of a set of contextually relevant elephants). As we will show in more detail in Dataset Construction below, the core semantic meaning of adjectives allows humans to draw inferences between expressions that contain adjectives, both about the basic meaning of adjectives and also about other related phenomena, such as when additional degree operators have been applied to adjectives or when comparative and superlative constructions are at play. What is interesting for the current purpose is that although two different adjectives, for instance *straight* and *warm*, may not necessarily share similar contexts of occurrences due to the fact that these words are describing very different attributes of objects in the world, but our understanding of certain abstract aspects of their meaning, such as how they apply to an argument and the possible resulting inferences we can draw, can actually be similar across different lexical items.

*Corresponding author: Nai Ding

[1]We mainly focus on gradable adjectives in our work, which we will refer to in the text simply as adjectives.

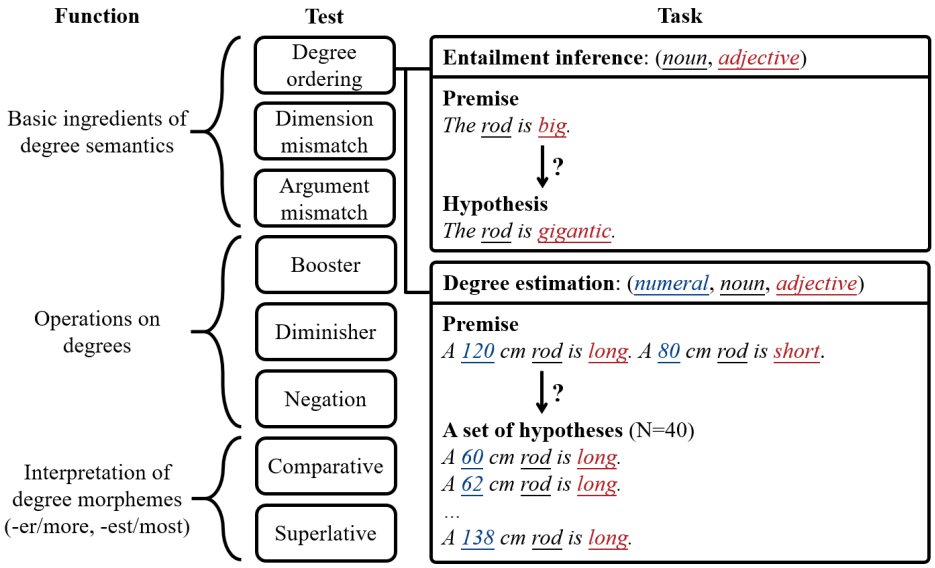

Figure 1: The ASP dataset is divided into 8 tests targeting at different aspects of adjective interpretation. In the examples, words that are filled into the template are underlined.

Test	Sentence	
Degree ordering	The rod is big. The rod is gigantic.	E↑ N↓
Dimension mismatch	The rod is big. The rod is excellent.	N↑ N↓
Argument mismatch	The rod is big. The pole is big.	N↑ N↓
Booster	The rod is big. The rod is very big.	E↑ N↓
Diminisher	The rod is big. The rod is relatively big.	N↑ E↓
Negation	The rod is big. (relative adj.) The rod is not small.	N↑ E↓
Comparative	The rod is bigger than the pole. The rod is big. (relative adj.) The pole is not big.	N↓ N↓
Superlative	The rod is the longest rod in the world. The rod is the longest rod I have ever seen.	N↑ E↓

Table 1: Examples for the entailment inference task. The arrow stands for the inference direction from the premise to the hypothesis. E stands for entailment, and N stands for not-entailment. *Negation* and *Comparative* tests vary for different class of adjectives. See detailed construction templates in Appendix Table 3.

One of the main goals of the current study is to investigate whether language models, representing word meaning only via its context, can capture human's understanding of such words. Based on the degree semantics analysis of adjectives, we build a fine-grained and theoretical-motivated dataset, to probe the models' understanding on the degree semantic information of adjectives.

The diagnostic dataset we build, referred to as the Adjective Scale Probe (ASP), is formulated using the Natural Language Inference (NLI) task, which requires models to identify the entailment relation between a pair of sentences (Dagan, Glickman, and Magnini 2006; Bowman et al. 2015). NLI is a basic task for evaluation and can be flexibly adapted to test specific aspects of language comprehension (Poliak 2020). In recent years, a growing body of NLI-style datasets aim to evaluate specific linguistic capabilities such as numerical reasoning (Ravichander et al. 2019) and pragmatic inference (Jeretic et al. 2020), and to analyze superficial heuristics learned by language models (McCoy, Pavlick, and Linzen 2019; Dev et al. 2020). Here, we create ASP to systematically probe the adjective understanding of language models. ASP contains 8 tests (see Dataset Construction), separately evaluate 3 main aspects of adjective interpretation, i.e., the basic ingredients of degree semantics, operations on degrees, and degree morphemes in comparative and superlative constructions. We semi-automatically create the ASP dataset based on syntactically simple sentence templates and a diverse set of vocabulary items, and apply several measures to ensure the plausibility of the combination between constituents.

We apply the ASP dataset to evaluate whether transformer-based language models can understand adjectives. Since the models all perform poorly on the ASP, we also fine-tune models on a subset of the ASP. It is shown that models fine-tuned on the ASP can generalize well to untrained adjectives and untrained tests. The main contributions of our study are: (i) constructing a theoretically motivated and fine-grained NLI diagnostic dataset to test models' understanding on the degree semantics of adjectives. (ii) demonstrating some language models can encode degree semantic information of adjectives, but need specific training data to learn how to apply such information to solve the inference tasks in current study.

Dataset Construction

Tasks For each test illustrated in Figure 1 (see the column *Test*), we design two tasks, i.e., a regular entailment

Test	Premise	Hypothesis	Entailment range
Degree ordering	A 120 cm rod is long. A 80 cm rod is short.	A α cm rod is long. A α cm rod is short.	$\alpha \geq 120$ - δ $\alpha \leq 80 + \delta$
Dimension mismatch	A 120 cm rod is straight. A 80 cm rod is bent.	A α cm rod is straight. A α cm rod is bent.	$\emptyset$ ($\alpha \neq 120$) $\emptyset$ ($\alpha \neq 80$)
Argument mismatch	A 120 cm rod is long. A 80 cm rod is short.	A α cm pole is long. A α cm pole is short.	$\emptyset$ $\emptyset$
Booster	A 120 cm rod is long. A 80 cm rod is short.	A α cm rod is very long. A α cm rod is very short.	$\alpha \geq 120 + \delta$ $\alpha \leq 80 - \delta$
Diminisher	A 120 cm rod is long. A 80 cm rod is short.	A α cm rod is relatively long. A α cm rod is relatively short.	$\alpha \geq 120$ - δ $\alpha \leq 80 + \delta$
Negation	A 120 cm rod is long. A 80 cm rod is short.	A α cm rod is not long. A α cm rod is not short.	$\alpha \leq 80 + \delta$ $\alpha \geq 120 - \delta$
Comparative	Rods longer than 120 cm is long. Rods shorter than 80 cm is short.	A α cm rod is long. A α cm rod is short.	$\alpha \geq 120$ $\alpha \leq 80$
Superlative	The longest rod in the world is 138 cm. The shortest rod in the world is 60 cm.	A α cm rod is long. A α cm rod is short.	$\alpha \geq 138$ - δ $\alpha \leq 60 + \delta$

Table 2: Examples for the degree estimation task. The *entailment range* column shows the entailment condition of α for the sentence pair. In these examples, $\alpha \in$ [60,138], with a fixed interval of 2. $\emptyset$ means the entailment relation is not satisfied for any numeral. δ is a positive value fitted by human annotation (see Appendix B). See detailed construction templates in Appendix Table 4.

inference task and a degree estimation task. Both are versions of NLI tasks that require models to identify whether the premise could or could not entail a hypothesis. All premises and hypotheses are constructed using templates. For the entailment inference task, the critical information in the premises and hypotheses rests in the relationship between the noun and adjective used in each sentence, and a large number of nouns and adjectives are used to construct the tests.

The degree estimation task also tests models' understanding of pairings of nouns and adjectives, but critically we also add numeral information to the tests in order to more precisely target models' understanding of the crucial concept of comparison threshold in degree semantics. This task focuses on a relatively small number of adjectives that can be mapped to physical dimensions that are easy to quantify numerically (i.e. length, mass, price, and temperature). For the degree estimation task, each premise contains an antonym pair, and each premise is tested with 40 hypotheses that only differed in the numeral, which is sampled from given range with fixed interval (see an example in Figure 1). Models are required to judge the entailment relation between the premise and each hypothesis from the set of hypotheses.

In total, we generate 8K NLI sample for the entailment inference task (1K for each test), and 16K NLI samples for the degree estimation task (2K for each test). We apply several measures to ensure the plausibility of samples (see Appendix A). See examples of each test in Tables 1 and 2. In the following, we detail how the premises and hypotheses are constructed for each test.

Basic Ingredients of Degree Semantics According to degree semantics, an adjective scale is a triplet of the following parameters (Kennedy 2007): a set of degrees, an ordering relation, and a dimension of measurement. The adjective maps its argument to a degree on the measurement scale, ordered with respect to other degree(s) on the same scale. Targeted at these parameters, we construct 3 tests, i.e., *Degree ordering*, *Dimension mismatch*, and *Argument mismatch*, to evaluate models' understanding of these basic ingredients of degree semantics. The *Degree ordering* test targets the degree ordering between a pair of adjectives that can be mapped to the same measurement scale. For this test, the entailment inference task (Table 1) looks at the asymmetrical entailment relation between two adjectives, one of which expresses a greater degree than the other (e.g. *gigantic* vs. *big*). When the two adjectives are mapped to the same noun phrase subject, the assertion with the stronger degree (*gigantic*) entails the assertion with the weaker degree (*big*), but not vice versa. The degree estimation task (Table 2) makes use of pairs of antonyms, which render a reverse ordering of the degrees on the same measurement scale. For instance, the semantics of a positive adjective *long* specifies the length of an object is larger than a threshold degree (upper open), whereas its antonym *short* reverses the relation (lower open). The *Dimension mismatch* test examines whether the models understand that there is no ordering relation between degrees that come from distinct measurement dimensions (i.e. distinct scales). For instance, for a sentence "*The rod is big/excellent*" (Table 1), there is no relationship between the size of a *rod* and the goodness of a *rod*, even though the adjective *excellent* expresses a large degree on the goodness-scale, parallel to the status of *gigantic* on a size scale. Similarly, for the example in Table 2, the length of a *rod* is not informative for evaluating the straightness of the *rod*. The *Argument mismatch* test examines whether the models understand that the evaluation of an adjective is argument-dependent. For the example in Table 1, the evaluation of *big* in the context of the *rod* is independent from the evaluation of *big* in the context of the *pole*. Similarly, for the example in Table 2, the length evaluation of the *rod* is irrelevant for the length evaluation of the *pole*.

Model	Entailment inference								Degree estimation							
	Ingredient			Operation			Morpheme		Ingredient			Operation			Morpheme	
	Ord.	Dim.	Arg.	Bo.	Di.	Ne.	Com.	Sup.	Ord.	Dim.	Arg.	Bo.	Di.	Ne.	Com.	Sup.
BERT-base	56.9	89.9	60.5	52.7	30.8	70.6	44.7	44.2	68.8	94.9	94.0	75.3	58.2	45.3	58.0	52.9
BERT-large	53.7	87.7	67.0	51.4	32.7	**71.1**	43.8	43.1	70.4	94.9	97.8	76.6	62.8	46.6	72.7	61.5
DeBERTa-base	56.2	94.0	81.6	**59.8**	10.4	68.0	38.2	31.2	**78.8**	83.1	97.0	74.2	**68.3**	**70.4**	76.1	**72.5**
DeBERTa-large	**59.4**	96.1	85.6	55.8	3.8	67.9	47.6	55.2	74.1	95.7	99.4	**78.7**	61.7	69.4	**76.7**	59.9
T0 3B	52.2	**97.0**	85.2	50.7	48.3	43.4	**57.1**	50.6	72.6	**95.8**	**99.6**	77.9	59.1	44.2	74.8	56.8
T0 pp	57.2	94.6	**86.4**	50.1	**50.2**	64.7	55.6	**56.7**	70.1	73.9	76.1	71.4	64.6	59.0	61.8	63.4
Chance level	50.0	66.6	66.6	50.0	50.0	41.7	55.6	50.0	55.8	50.0	50.0	60.8	52.1	47.9	58.3	52.1
Majority baseline	50.0	100.0	100.0	50.0	50.0	75.0	66.7	50.0	67.5	100.0	100.0	81.3	56.3	44.3	75.0	56.3

Table 3: Performance of zero-shot models and NLI models on the ASP. The best model performance for each test is highlighted in bold. The majority baseline is calculated by predicting all samples as entailment or not entailment. The chance level baseline is calculated by randomly guessing the prediction. Ord.: *Degree ordering*, Dim.: *Dimension mismatch*, Arg.: *Argument mismatch*, Bo.: *Booster*, Di.: *Diminisher*, Ne.: *Negation*, Com.: *Comparative*, Sup.: *Superlative*.

Operations on Degrees The relation between a degree that an object instantiates and a threshold can be further manipulated through degree modifiers. For example, a booster, such as *very*, *extremely*, etc., "boosts the meaning of a property upwards from an assumed norm" (Quirk et al. 1987). On the other hand, a diminisher adverb (e.g., *relatively*, *mildly*) would weaken the strength of the adjective they combine with (Paradis 2008). We conduct 3 tests to examine whether the models understand that one can apply various operations to the degrees on a scale. We focus on the *Booster* and *Diminisher* adverbs, as well as the *Negation* operator. For both the entailment inference and degree estimation tasks, each premise-hypothesis pair shares the same adjective and the hypothesis is different from its premise by an operator (a booster, diminisher or negation). The only slight variation is that, for the hypothesis in the *Negation*-entailment inference test (Table 1), instead of applying negation to the same adjective used in the premise, we use "*not* + antonym" to avoid the simple contradiction-bias of *not* (Gururangan et al. 2018; He, Zha, and Wang 2019). Notably, the entailment patterns based on the "*not* + antonym" form also allow us to probe the context-sensitivity of various classes of adjectives , i.e., absolute adjective (*bent*, *straight*) vs. relative adjective (*big*, *small*) (see a discussion of different classes of adjectives in (Pinkal 1995; Rotstein and Winter 2004; Kennedy 2007; Toledo and Sassoon 2011; Solt 2012)).

Degree Morphemes The tests discussed above mainly evaluate the semantics of bare adjectives. In English at least, explicit degree morphemes (e.g. *–er*/*-est*, or *more*/*most*) are available for comparative and superlative constructions. We examine the models' understanding on the explicit degree morphemes via the *Comparative* and *Superlative* tests. For the entailment inference task, *Comparative* and *Superlative* tests are designed to probe whether the models understand the context-dependency of comparison threshold and domain restriction. The entailment patterns of the comparative form also reflect the nuanced differences between different classes of adjectives in a more fine-grained way (Kennedy 2007), i.e., relative adjective (*big*, *long*) vs minimum absolute adjective (*bent*, *wet*) vs. maximum absolute adjective (*straight*, *safe*). For the degree estimation task, via setting the explicit threshold for a *Comparative* or *Superlative* evaluation, we test the models' understanding on the basic meaning of degree morphemes.

Model Performance on the ASP

Experimental Setup

We tested NLI models and zero-shot models on the ASP dataset. NLI models were the pre-trained language models fine-tuned on MNLI (Williams, Nangia, and Bowman 2018), while zero-shot models were only pre-trained. The NLI models included both base and large versions of BERT (Devlin et al. 2019) and DeBERTa-v3 (DeBERTa in short. He, Gao, and Chen 2021). The zero-shot models included T0 3B and T0 pp, both were variants of T0 that were not trained to perform the NLI task (Sanh et al. 2021). See detailed fine-tuning (for the NLI models) and inference procedures in Appendix C.

To evaluate a model on the ASP, we collapsed the 3-label classification of NLI task to 2 labels, with the neutral and contradiction labels collapsed to a single not-entailment label (McCoy, Pavlick, and Linzen 2019). For the entailment inference task, we used accuracy as the evaluation metric. For the degree estimation task with human annotation, we used the sigmoid function to fit the human results, and took the median (proportion=50%) of fitted curves as the threshold of each test type (i.e. δ in Table 2). We used the threshold to calculate the accuracy of models. We visualized the model predictions on the degree estimation task following the similar procedure of human results.

Result

The model performance on the ASP is shown in Table 3. For the entailment inference task, the accuracy of the best performing model (highlighted in bold) was never more than 10% above the majority baseline. All models failed to correctly distinguish the degree difference between lexical pairs (*Degree ordering* test), and adjective phrases (*Booster* and *Diminisher* test). For example, all models tended to judge that "*The rod is big*" entails "*The rod is gigantic*". Similarly, all models tended to judge that "*The rod is big*" entails "*The rod is very big*". The performance was relatively high on

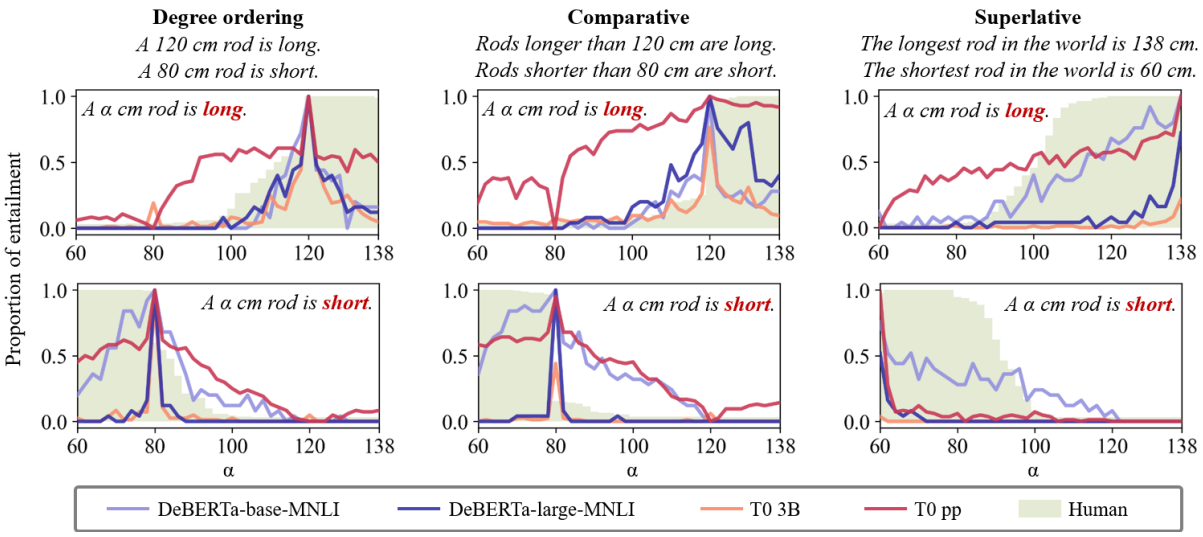

Figure 2: Model performance on the *Degree ordering*, *Comparative*, and *Superlative* tests of the degree estimation task.

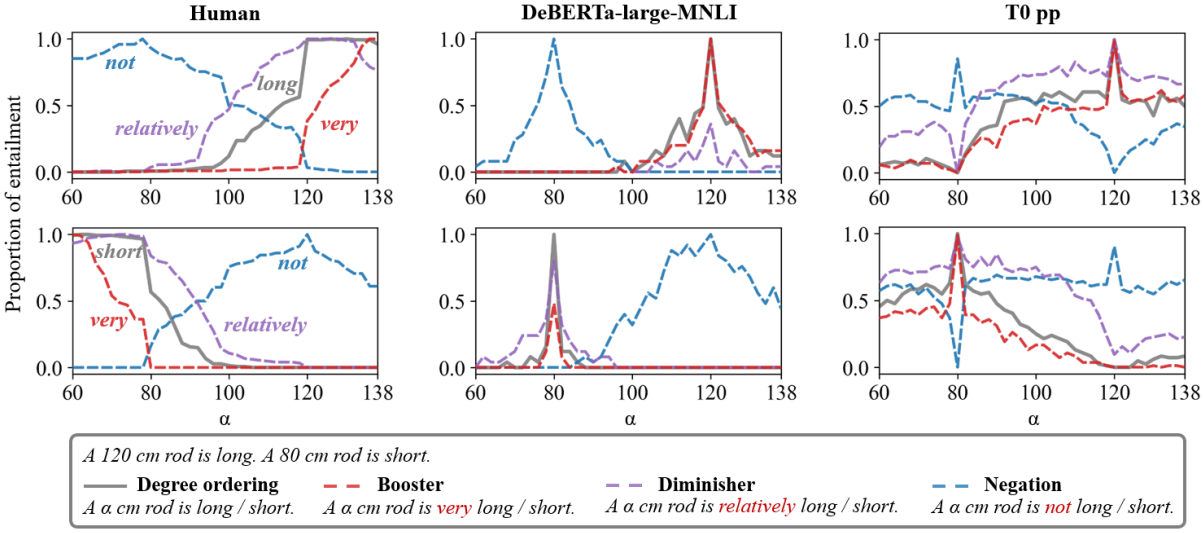

Figure 3: Model performance on the *Booster*, *Diminisher* and *Negation* tests of the degree estimation task.

the *Dimension mismatch* and *Argument mismatch* tests. In the *Negation* test, models tended to judge that the negation form of an adjective was always equivalent to its antonym, i.e., "*the rod is not small*" entails "*the rod is big*", regardless of the semantic class of adjectives. Furthermore, all models were confused about the context-sensitivity of threshold (*Comparative* test) and domain restriction (*Superlative* test).

For the degree estimation task, as shown in Table 3, the models performed poorly on most of the tests except for the *Dimension mismatch* and *Argument mismatch* tests. We visualized the behavior of a few tests in Figures 2 and 3, with the full set of results presented in Appendix Figures 2-4. In Figure 2 we presented the results from 3 tests: *Degree ordering*, *Comparative* and *Superlative*. For human annotators, when presented with a pair of antonyms such as *long-short*, they used the numeral information in the premise to help gauge the range on a scale that holds true for the adjectival predicate. For instance, a premise sentence "*A 120 cm rod is long. A 80 cm rod is short*" would indicate to a human annotator that any length above 120 cm was *long* and there was also some (monotonically increasing) likelihood that an area on the scale prior to 120 cm, i.e. 100-120 cm in Figure 2 top left panel, the *rod* could also be considered as *long*. Their understanding of the antonym word short was reversed (Figure 2 bottom left). Across all the tests presented in Figure 2, human results showed a clear understanding of degree semantics. In contrast, the model predictions, in comparison to the human results, revealed poor performance. For the most parts, the models almost exclusively favored the single numerical value provided in the premise, failing to draw inferences about other values on the scale. T0 pp and DeBERTa-base, although still far from satisfactory, had slightly better

Model	Entailment inference								Degree estimation							
	Ingredient (training)			Operation			Morpheme		Ingredient (training)			Operation			Morpheme	
	Ord.	Dim.	Arg.	Bo.	Di.	Ne.	Com.	Sup.	Ord.	Dim.	Arg.	Bo.	Di.	Ne.	Com.	Sup.
BERT-base	96.4	94.6	98.5	**54.7**	47.0	46.6	55.1	49.9	62.5	**97.6**	99.9	69.9	52.9	46.7	66.6	50.6
BERT-large	**98.0**	94.4	98.8	48.9	50.2	26.6	64.8	**59.9**	45.8	81.5	99.1	36.7	46.6	49.5	61.5	53.6
DeBERTa-base	97.0	**98.9**	**99.9**	50.4	50.0	23.5	**66.7**	50.0	84.1	98.1	**100.0**	70.8	88.1	26.9	**72.2**	**77.2**
DeBERTa-large	97.2	98.4	98.9	46.6	**50.7**	**40.1**	66.6	56.2	**87.1**	95.8	99.2	**77.6**	**92.4**	**59.4**	71.3	68.6
Chance level	50.0	66.6	66.6	50.0	50.0	41.7	55.6	50.0	55.8	50.0	50.0	60.8	52.1	47.9	58.3	52.1
Majority baseline	50.0	100.0	100.0	50.0	50.0	75.0	66.7	50.0	67.5	100.0	100.0	81.3	56.3	43.8	75.0	56.3

Table 4: Performance of models fine-tuned on a subset of the ASP. The best model performance for each test is highlighted in bold.

performance relative to other models, sharing some similarity with the human results.

In Figure 3 we presented the visualization for the 3 Degree operation tests: *Booster*, *Diminisher* and *Negation*. For a human annotator, not only understood what is an appropriate length distribution for a *long rod* (the grey distribution in the left panel in Figure 3), he/she also understood a booster like *very* shifts the distribution upwards on the scale, a diminisher like *relatively* shifts the distribution downwards, and a negation operator would take the complement set of degrees on the scale. For the models, only the T0 pp showed a somewhat promising trend similar to the human results. But by and large, the models failed to match human results.

The evaluation result indicated that NLI models and zero-shot models performed poorly on our ASP dataset. T0 pp and DeBERTa-base showed some human-like behavior in a few degree estimation tests. We further measured the correlation (R=0.20, p=0.17) between the performance on the degree estimation task and numerical reasoning ability (Appendix Figure 5), and excluded the possibility that numerical reasoning ability alone led to the models' poor performance on the degree estimation task.

Fine-tuning on the ASP

Experimental Setup

Though models performed badly on the ASP, it was possible that the models maintained some degree semantic information of adjectives but failed to apply such information to solve NLI tasks. Therefore, in the following, we fine-tuned both base and large versions of BERT and DeBERTa, using a small subset of the ASP and tested whether the fine-tuning effect could transfer to untrained adjectives and ASP tests.

Specifically, we used samples from the *Degree ordering*, *Dimension mismatch*, and *Argument mismatch* tests for fine-tuning, all of which were designed to probe the basic ingredients of degree semantics. The fine-tuning was done 4 times, involving both the entailment inference and degree estimation tasks. For the entailment inference task, we split the adjective vocabulary into training/testing set before data generation. Each time, we used 50% of the adjectives to construct the training set, and left the remaining half of the adjectives for testing. For the degree estimation task, each time, we used 3 physical dimensions for training, e.g., length, mass, and price, and the remaining dimension, e.g., temperature, was used for testing. Additionally, the hypothesizes we used for training only covered a subset of the scale, in particular the area on the scale with $\alpha < 80$ and $\alpha > 120$.

In sum, we used 6000 samples (1500 samples for the entailment inference task, 4500 samples for the degree estimation task) for fine-tuning. The label distribution of training set was balanced. The fine-tuning parameters on the ASP were shown in Appendix Table 2. We reported below the evaluation results on the testing sets, averaged across 4 fine-tuning procedures.

Result

The model performance on the testing sets is shown in Table 4. For the entailment inference task, the models achieved $\sim$100% accuracy on the untrained lexical items in the *Degree ordering* test, while still failed on the other tests that were withheld from training. For the degree estimation task, by learning the entailment patterns on the basic ingredients of degree semantics, models could generalize the capability to some other untrained tests, such as the *Diminisher* and *Superlative* tests.

In Figure 4, we visualized the results for a few tests of the degree estimation task. In the *Degree ordering* test, although the range of $80 < \alpha < 120$ was withheld from training, DeBERTa models showed a gradual increase of entailed hypotheses over this range, similar to human results. Similarly, although the *Booster*, *Diminisher* and *Negation* tests were withheld from training, DeBERTa-large showed behavior similar to human results. In Appendix Figures 6 and 7 we presented the visualization results for other tests and models. In general, DeBERTa models performed better to generalize from training sets to untrained tests, whereas BERT models performed more poorly.

To summarize, the results showed improvements of performance after fine-tuning on a subset of the ASP. Although the models were still not performing at ceiling level, it was impressive that they could transfer the learning outcome from training to untrained lexical items and tests. This was indicative of the possibility that the pre-trained models to some extent had access to the abstract degree semantics information, but fine-tuning on MNLI might not provide sufficient signals that underscored the need to deploy such information. As a result, without fine-tuning on the ASP, the models did not know how to apply the knowledge of degree semantics to the specific inference tasks in the current study. Yet another remaining possibility was that the models developed some superficial heuristics during fine-tuning. We

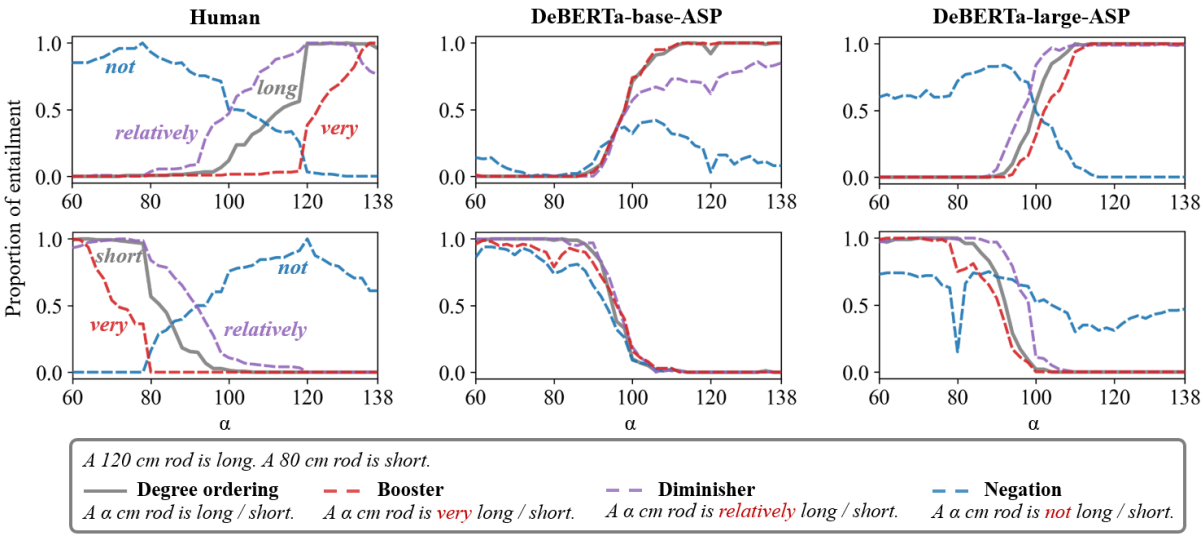

Figure 4: Model performance on the *Booster*, *Diminisher*, and *Negation* tests of the degree estimation task.

further conducted control experiments to rule out this possibility (see Appendix D). Additionally, we manipulated the numerals in the premise to evaluate the robustness of models (see Appendix E), and the results were similar to the current results.

Related Work

Current language models are data driven, therefore constructing informative datasets is critical to train and evaluate models. Recently, a large number of datasets have been proposed. Some contains more challenging samples (Nie et al. 2020) and some are diagnostic datasets targeted at evaluating particular linguistic capabilities (Ravichander et al. 2019; Richardson et al. 2020; Saha, Nie, and Bansal 2020; Vashishtha et al. 2020; Kober, Bijl de Vroe, and Steedman 2019). Our study contributes to this body of literature in theoretically informative ways. The ASP is a fine-grained diagnostic dataset that comprehensively tests language models' understanding of adjectives, a major class of words in natural language. Our tests are grounded in the formal theory (i.e., degree semantics) of adjective semantics, borrowing the insights of formal semantics to help better interpret the internal representations of language models. Some previous studies have also analyzed the properties of the word embeddings of adjectives under the framework of degree semantics: Garí Soler and Apidianaki (2020) find that a diagnostic classifier can decode the intensity of adjectives from word embeddings of BERT. Samir, Beekhuizen, and Stevenson (2021) classify the extremeness of adjectives based on word2vec. Both of these studies aim at distinguishing different types of adjective classes. The current study focuses on whether models can correctly capture the entailment relations between sentences, a task that is more closely related to what humans do when they understand sentence meaning. In addition to the traditional entailment inference task, which only elicits a categorical response from the models, we also develop a new type of entailment inference task, the degree estimation task, which quantifies models' responses on a numeral scale.

One of the interesting findings of our study is that models fine-tuned on a subset of the ASP dataset perform much better than models that are only fine-tuned on MNLI. This highlights the need to be cautious when drawing conclusions based on the failure of a model. When models fail to perform a task, the problem could arise from the inherent inadequacy of the internal structure of the model or it could be due to insufficient training signals. At least for the current case, our study suggests that pre-trained language models can encode to some extent the degree semantic information of adjectives. But the models need to be trained to understand how to apply such information. This is consistent with previous findings that suggest models may have learned superficial heuristics from the large-scale NLI datasets (Gururangan et al. 2018; Naik et al. 2018; McCoy, Pavlick, and Linzen 2019). Models fine-tuned on NLI datasets may fail on linguistically more sophisticated tests, but capability-specific datasets can potentially improve performance on these evaluations (Liu, Schwartz, and Smith 2019; Wang et al. 2022).

Conclusion

In summary, based on the degree semantics analysis of adjectives, we create the ASP dataset to comprehensively and quantitatively evaluate the understanding of adjectives of language models. The current study shows that although the state-of-the-art transformer-based language models reach human performance in popular datasets such as MNLI, they fail to precisely understand the meaning of adjectives. By fine-tuning pre-trained models on a subset of the ASP, models can generalize the learning outcome to untrained lexical items and tests, indicative of the possibility that pre-trained models can encode (at least some) formal semantic information of adjectives. But the models need the specific training data to learn how to apply such information to solve the inference tasks in the current study.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. This research is supported by the National Key Research and Development Program of China (No. 2021ZD0204105). Ming Xiang is supported by the University of Chicago Humanities Division Council.

References

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Online: Association for Computational Linguistics.

Bisk, Y.; Holtzman, A.; Thomason, J.; Andreas, J.; Bengio, Y.; Chai, J.; Lapata, M.; Lazaridou, A.; May, J.; Nisnevich, A.; Pinto, N.; and Turian, J. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8718–8735. Online: Association for Computational Linguistics.

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

Cresswell, M. 1976. The semantics of degree. In PARTEE, B. H., ed., *Montague Grammar*, 261–292. Academic Press. ISBN 978-0-12-545850-4.

Dagan, I.; Glickman, O.; and Magnini, B. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, 177–190. Berlin, Heidelberg: Springer Berlin Heidelberg.

Dev, S.; Li, T.; Phillips, J. M.; and Srikumar, V. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05): 7659–7666.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Garí Soler, A.; and Apidianaki, M. 2020. BERT Knows Punta Cana is not just beautiful, it's gorgeous: Ranking Scalar Adjectives with Contextualised Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7371–7385. Online: Association for Computational Linguistics.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 107–112. New Orleans, Louisiana: Association for Computational Linguistics.

He, H.; Zha, S.; and Wang, H. 2019. Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, 132–142. Hong Kong, China: Association for Computational Linguistics.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Heim, I. 2000. Degree operators and scope. In *Semantics and linguistic theory*, volume 10, 40–64.

Jeretic, P.; Warstadt, A.; Bhooshan, S.; and Williams, A. 2020. Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8690–8705. Online: Association for Computational Linguistics.

Kennedy, C. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, 30(1): 1–45.

Kennedy, C.; and McNally, L. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, 81(2): 345–381.

Kober, T.; Bijl de Vroe, S.; and Steedman, M. 2019. Temporal and Aspectual Entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, 103–119. Gothenburg, Sweden: Association for Computational Linguistics.

Lin, J.; Zou, J.; and Ding, N. 2021. Using Adversarial Attacks to Reveal the Statistical Bias in Machine Reading Comprehension Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 333–342. Online: Association for Computational Linguistics.

Liu, N. F.; Schwartz, R.; and Smith, N. A. 2019. Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. 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)*, 2171–2179. Minneapolis, Minnesota: Association for Computational Linguistics.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448. Florence, Italy: Association for Computational Linguistics.

Miaschi, A.; Brunato, D.; Dell'Orletta, F.; and Venturi, G. 2020. Linguistic Profiling of a Neural Language Model. In

Proceedings of the 28th International Conference on Computational Linguistics, 745–756. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Naik, A.; Ravichander, A.; Sadeh, N.; Rose, C.; and Neubig, G. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2340–2353. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901. Online: Association for Computational Linguistics.

Paradis, C. 2008. Configurations, construals and change: expressions of DEGREE. *English Language and Linguistics*, 12(2): 317–343.

Pinkal, M., ed. 1995. *Logic and Lexicon*. Springer Dordrecht.

Poliak, A. 2020. A survey on Recognizing Textual Entailment as an NLP Evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 92–109. Online: Association for Computational Linguistics.

Quirk, R.; Greenbaum, S.; Leech, G.; and Svartoik, J. 1987. *A comprehensive grammar of the English language*. London ; New York : Longman, 1985.

Ravichander, A.; Naik, A.; Rose, C.; and Hovy, E. 2019. EQUATE: A Benchmark Evaluation Framework for Quantitative Reasoning in Natural Language Inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 349–361. Hong Kong, China: Association for Computational Linguistics.

Richardson, K.; Hu, H.; Moss, L.; and Sabharwal, A. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8713–8721.

Rotstein, C.; and Winter, Y. 2004. Total Adjectives vs. Partial Adjectives: Scale Structure and Higher-Order Modifiers. 12: 259–288.

Saha, S.; Nie, Y.; and Bansal, M. 2020. ConjNLI: Natural Language Inference Over Conjunctive Sentences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8240–8252. Online: Association for Computational Linguistics.

Samir, F.; Beekhuizen, B.; and Stevenson, S. 2021. A Formidable Ability: Detecting Adjectival Extremeness with DSMs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4112–4125. Online: Association for Computational Linguistics.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Solt, S. 2012. Comparison to arbitrary standards. In *Proceedings of Sinn und Bedeutung*, volume 16, 557–570.

Stechow, A. v. 1984. Comparing semantic theories of comparison. *Journal of Semantics*, 3(1-2): 1–77.

Toledo, A.; and Sassoon, G. W. 2011. Absolute vs. relative adjectives-variance within vs. between individuals. *Semantics and linguistic theory*, 21: 135–154.

Vashishtha, S.; Poliak, A.; Lal, Y. K.; Van Durme, B.; and White, A. S. 2020. Temporal Reasoning in Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4070–4078. Online: Association for Computational Linguistics.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. 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)*, 2153–2162. Hong Kong, China: Association for Computational Linguistics.

Wang, X.; Liu, B.; Xu, F.; Long, B.; Tang, S.; and Wu, L. 2022. Feeding What You Need by Understanding What You Learned. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5858–5874. Dublin, Ireland: Association for Computational Linguistics.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Yenicelik, D.; Schmidt, F.; and Kilcher, Y. 2020. How does BERT capture semantics? A closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 156–162. Online: Association for Computational Linguistics.