ADEPT: An Adjective-Dependent Plausibility Task

Ali Emami[1], **Ian Porada**[1], **Alexandra Olteanu**[2],
Kaheer Suleman[2], **Adam Trischler**[2], and **Jackie Chi Kit Cheung**[1]

[1]Mila, McGill University
[2]Microsoft Research Montréal
{*ali.emami, ian.porada*}*@mail.mcgill.ca*
{*alexandra.olteanu, adam.trischler, kasulema*}*@microsoft.com*
jcheung@cs.mcgill.ca

Abstract

A *false contract* is *more likely* to be rejected than a *contract* is, yet a *false key* is *less likely* than a *key* to open doors. While correctly interpreting and assessing the effects of such adjective-noun pairs (e.g., *false key*) on the plausibility of given events (e.g., *opening doors*) underpins many natural language understanding tasks, doing so often requires a significant degree of world knowledge and common-sense reasoning. We introduce ADEPT – a large-scale semantic plausibility task consisting of over 16 thousand sentences that are paired with slightly modified versions obtained by adding an adjective to a noun. Overall, we find that while the task appears easier for human judges (85% accuracy), it proves more difficult for transformer-based models like RoBERTa (71% accuracy). Our experiments also show that neither the adjective itself nor its taxonomic class suffice in determining the correct plausibility judgement, emphasizing the importance of endowing automatic natural language understanding systems with more context sensitivity and common-sense reasoning.

1 Introduction

Discerning the varying effects of adjectival modifiers on the reading of a sentence is critical in a variety of tasks involving natural language understanding. Consider the following examples:

(1) a. A [dead] monkey turns on a light switch.
b. A [dead] leg has one foot.
c. A [dead] leaf falls from a tree in autumn.

The reading of these sentences with and without the modifier *dead* is notably different. The plausibility judgement of the event where a monkey turns on a light switch decreases when the adjectival modifier *dead* is added, while in the 1b or 1c examples, adding the same modifier leads to no change or an increase in event plausibility, respectively.

This observation has important ramifications for many NLP applications like information extraction (IE) and recognizing textual entailment (RTE), where solutions have often relied on normative rules that group the effects of adjectives according to either the adjective or its taxonomic class (McNally and Boleda, 2004; Amoia and Gardent, 2007; McCrae et al., 2014). These taxonomies distinguish adjectives like *false, dead, alleged* (non-subsective) from others like *red, large, or valid* (subsective).

Specifically, while the 1a example may influence systems to adopt the rule that adding a non-subsective adjective like *dead* to a noun leads to a decrease in plausibility, the other examples suggest a conflicting rule. Distinguishing the effects of different adjectives (beyond just their denotation) may thus require common-sense and world knowledge.

Powerful, massively pre-trained language models (LMs) have pushed the performance on various natural language understanding benchmarks to impressive figures; transformer architectures including BERT and RoBERTa are believed to perform at near human-level performance on a number of Natural Language Inference (NLI) tasks (Liu et al., 2019), while the recently proposed DeBERTa, which builds upon the former two architectures, performs at state-of-the-art on MNLI, RTE, QNLI and WNLI (He et al., 2020). It is however unclear whether the complex effects of the classes of modifiers exampled above are captured by the competing models given their sparsity in both the corpora and existing NLI benchmarks.

To examine the ability of LMs to capture and distinguish the effects of adjectives on events plausibility, we present a challenge task formulated as a plausibility classification problem consisting of sentence pairs with and without inserting possible adjectives. We do so to understand the strengths and weaknesses of LMs that have led to state-of-the-art performance in downstream NLI-tasks. Ta-

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 7117–7128
August 1–6, 2021. ©2021 Association for Computational Linguistics

ADEPT instance	Inserted Modifier (Taxonomic Class)	Plausibility Change
(1a): A [false] key opens doors.	False (NS)	Less Likely
(1b): A [false] statement is a lie.	False (NS)	Necessarily True
(1c): A [false] alarm causes danger.	False (NS)	More Likely
(2a) An [outstanding] year is made up of 365 days.	Outstanding (S)	Equally Likely
(2b) An [outstanding] coach pushes his players.	Outstanding (S)	More Likely
(2c) An [outstanding] professor waits for tenure.	Outstanding (S)	Less Likely
(3a) A [dead] monkey turns on a light switch.	Dead (NS)	Impossible
(3b) A [dead] leg has one foot.	Dead (NS)	Equally Likely
(3c) A [dead] leaf falls from a tree in autumn.	Dead (NS)	More Likely
(4a) An [old] graveyard is for settings for horror movies.	Old (S)	More Likely
(4b) An [old] parrot lays and egg.	Old (S)	Less Likely
(4c) An [old] nun prays.	Old (S)	Equally Likely

Table 1: Examples of ADEPT instances, revealing the diverse effects on plausibility change due to different adjectival modifiers. The plausibility change depends more on the context of the sentence, and less on the modifier or its taxonomic class.

ble 1 illustrates the task with several examples. Our contributions are three-fold:

We introduce a novel plausibility task: Using automated mechanisms to extract, filter and construct natural sentences, we create ADEPT—a large human-labeled semantic plausibility task consisting of 16 thousand pairs of sentences that differ only by one adjective added to a noun, and designed to resist the statistical correlations that might underpin modern distributional lexical semantics.[1]

We show that transformer-based models are not yet *adept* at ADEPT: Our findings suggest performance gaps between humans and large language representation models on ADEPT, which appears to be in large part due to the models' insensitivity to context, indicating an important area for their improvement.

We show that the effect of adjectival modifiers on event plausibility is *context* dependent: We quantify the degree to which plausibility judgements vary for the same adjective and taxonomic class, finding that rules based only on the adjective or its denotation are insufficient when assessing the plausibility readings of events. For example, in our task, the non-subsective adjective like *dead* led to a decrease in events plausibility as frequently as it led to no change at all.

Building on prior work showing that normative rules are often broken for subsective adjectives (Pavlick and Callison-Burch, 2016), we investigate possible effects across *all* types of adjectives, beyond just the taxonomical categories. The scope of our analysis also goes beyond entailment effects, examining the effects on *plausibility*, which can be seen as both complimentary and even an extension to entailment tasks.

2 Background and Related Work

Taxonomy of adjectives: The taxonomic classification of adjectives into subsective and non-subsective categories originates from the works of Parsons (1970), Montague (1970), Clark (1970) & Kamp and Keenan (1975). Canonically, subsective adjectives modify a noun such that the extension of the adjective-noun pair is a subset of the extension of the noun alone (e.g., a *blue* fish is still a fish and a *loose* tooth is still a tooth). In contrast, non-subsective adjectives modify a noun such that the extension of the adjective-noun pair is not a subset of the noun's extension (e.g., a *former* president is not a president or an *alleged* criminal is not necessarily a criminal). Kamp and Partee (1995) further divided non-subsective adjectives in two categories: *privative* and *plain*. While when combined with nouns privative adjectives produce a disjoint set of entities from the original noun (e.g., *former* president does not fall under the class of presidents, making *former* a privative adjective), plain non-subsective adjectives do not guarantee this mutual exclusiveness (e.g., an *alleged criminal* may or may not be a *criminal*).

This classification scheme has been adopted for many NLP applications including IE and RTE (Amoia and Gardent, 2006, 2007; McCrae

[1] The corpus and the code to reproduce all of our experimental results are available at https://github.com/aemami1/ADEPT.

et al., 2014). For RTE, inference rules were developed according to whether the adjective was non-subsective or not. For IE, non-subsective adjectives were treated as special cases for extracting open IE relations (Angeli et al., 2015). We show that there is also a relation between an adjective and the plausibility of the rest of the clause, even for subsective adjectives. This has direct implications for the extraction of generalizable abstract knowledge that can be extracted from a corpus.

Aspects of this classification scheme have since been challenged, resulting in efforts to either expand on its definitions or abandon the taxonomy altogether. Del Pinal (2015) suggests that the meaning of certain nouns are only partially modified by non-subsective adjectives (e.g., only the functional features are modified), while Nayak et al. (2014) tackle the categorization problem with a statistical approach focused on the proportion of properties shared by the noun and the adjective noun pair. Even more recently, inference rules relying on the original taxonomy were observed not to be without exceptions; Pavlick and Callison-Burch (2016) used human annotators to highlight cases where the deletion of non-subsective adjectives from a sentence does not necessarily result in non-entailment.

These on-going examinations and revisions underpin a profound linguistic phenomenon of mutual dependence: while adjectives play a crucial role in the correct interpretation of a sentence context, the context words are just as instrumental in determining the effect of an adjective; resulting in a number of exceptions to taxonomically-based rules. Inspired by this, our work explores the broader question of how dependent the effect of *any* adjective (beyond their taxonomical class) is on the interpretation of a sentence. For this, we frame our exploration in terms of *changes in the plausibility* of events, which we believe it can be seen as an extension to entailment.

Recognizing Textual Entailment & Semantic Plausibility: The RTE Challenges were yearly sources of textual inference examples (Dagan et al., 2006) consisting of a three-way classification task with the inputs as sentence pairs $\{T,\ H\}$ with labels for *entailment*, *contradiction* or *unknown* (meaning T neither contradicts nor entails H). Variations of this task are also described in SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018).

The Johns Hopkins Ordinal Commonsense Inference (JOCI) task generalizes RTE to the problem of determining relative change in semantic plausibility on an ordinal 5-level Likert scale (from impossible to very likely) (Zhang et al., 2017). Other semantic plausibility datasets have collected judgments for the plausibility of single events (Wang et al., 2018b) and the plausibility of adjectives modifying a meronym (Mullenbach et al., 2019). Such plausibility tasks have often been solved using either data-driven methods (Huang and Luo, 2017; Sasaki et al., 2017) or pre-trained LMs (Radford et al., 2019).

Prior work has also collected human assessments of the plausibility of adjective-noun pairs (Lapata et al., 1999; Keller and Lapata, 2003; Zhang et al., 2019); however, this line of work specifically focuses on the plausibility of bi-grams without context, known as selectional preference.

3 The Task: ADEPT

We develop ADEPT, a semantic plausibility task that features over 16 thousand instances consisting of two sentences, where the second sentence differs from the first only by the inclusion of an adjectival modifier. Examples of these instances are in Table 1, where the inserted modifier is bracketed.

Formally, given the original sentence s and the modified sentence s', s' is identical to s except for the addition of an adjective a before the root noun of the original sentence. The task is to assess the plausibility difference in the reading of s' versus that of s. The possible plausibility ratings are:

1. **Impossible** — s' is improbable or illogical.
2. **Less likely** — s' is less likely than s.
3. **Equally likely** — s' is as plausible as s is.
4. **More likely** — s' is more likely than s.
5. **Necessarily true** — s' is true by necessity, including repetitive use of phrases or words that have similar meanings.

4 Dataset

To construct ADEPT, we scrape text samples from English Wikipedia and Common Crawl, extracting adjectival modifier-noun pairs that occur with high frequency. We then curated these pairs through a multi-stage pipeline to filter out extraction errors, typos, and inappropriate words, as well as oversample non-subsective adjectives which tend to be in the long-tail of a given corpora. We then use existing knowledge bases to find relevant predicates for the noun in the adjective-noun pair and compose natural sentences based on them. To an-

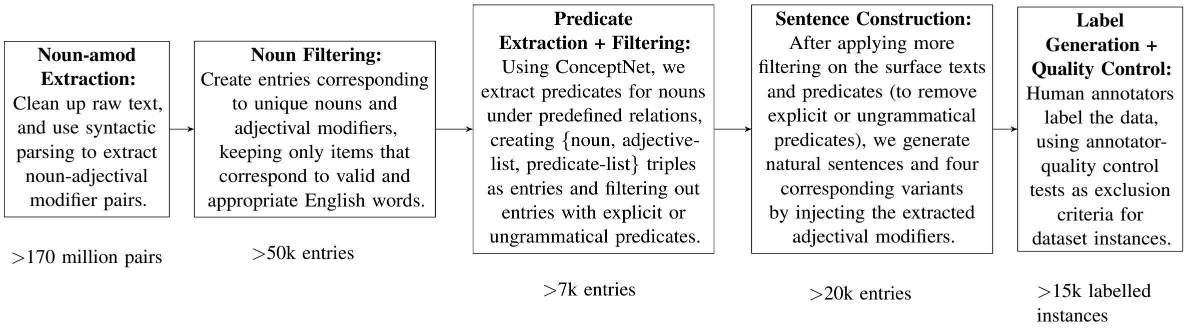
Noun-amod Extraction: Clean up raw text, and use syntactic parsing to extract noun-adjectival modifier pairs.
>170 million pairs
Noun Filtering: Create entries corresponding to unique nouns and adjectival modifiers, keeping only items that correspond to valid and appropriate English words.
>50k entries
Predicate Extraction + Filtering: Using ConceptNet, we extract predicates for nouns under predefined relations, creating {noun, adjective-list, predicate-list} triples as entries and filtering out entries with explicit or ungrammatical predicates.
>7k entries
Sentence Construction: After applying more filtering on the surface texts and predicates (to remove explicit or ungrammatical predicates), we generate natural sentences and four corresponding variants by injecting the extracted adjectival modifiers.
>20k entries
Label Generation + Quality Control: Human annotators label the data, using annotator-quality control tests as exclusion criteria for dataset instances.
>15k labelled instances

Figure 1: The overview of the data collection process for ADEPT.

notate the data, we provide human annotators with labelling instructions, while implementing quality control measures as exclusion criteria for final dataset instances.

4.1 Data Collection

We now detail the steps of our data collection process (see Figure 1 for an overview). Tables 2 and 3 provide examples of how each step contributes to the creation of an ADEPT instance.

Noun-amod extraction: In order to extract adjectival modifier and noun pairs, we use two dependency-parsed corpora: English Wikipedia, which we parse using the Stanza pipeline (Qi et al., 2020), and a subset of DepCC (Panchenko et al., 2018), an automatic parse of the Common Crawl corpus. After a preliminary examination of the modifier-noun pairs' quality, we kept only those pairs that occur at least 10 times in their respective corpus. This filtered out many pairs that appeared anomalous or atypical (e.g., *unwieldy potato*). We extracted 10 million pairs from English Wikipedia and 70 million pairs from Common Crawl.

Noun filtering: Using these pairs, we created dictionary items consisting of nouns—that co-occur with at least four different adjectival modifiers—along with their adjectival modifiers. This threshhold (as opposed to a higher one) allows us to both still find rare non-subsective adjectives to oversample at later steps, and avoid excessively reducing the number of extracted pairs.[2] We then filter out adjectives and nouns that e.g., are explicit, have offensive connotations using preset lists and automatic moderation tools (e.g., *profanity-filter* (Roman Inflianskas, 2020)). Finally, we ensured that both the nouns and adjectives are valid English words. This yielded slightly over 50 thousand noun-adjective dictionary items.

Predicate extraction: For the noun in each dictionary item, we use ConceptNet 5 (Liu and Singh, 2004) to find predicates under the relationships of *IsCapableOf*, *HasProperty*, *ReceivesAction*, *HasA*, and *UsedFor*. We restricted the predicates to these as they best characterize the *functional* features of a noun, which earlier studies found to be most sensitive to change according to the attaching modifier (Del Pinal, 2015). We also store the surface text—the sentence the ConceptNet annotator wrote to examplify a use of the predicate with the noun (e.g., for the noun "book," under the ConceptNet relation *IsCapableOf*, the predicate *include a table of contents* is found with the surface text: *A book can include a table of contents*).

Predicate Filtering + Scaling: After applying additional filtering to the surface texts and predicates (to remove explicit or ungrammatical predicates), we create triples containing a noun, a set of adjectival modifiers, and the predicates. This yielded over 7,000 triples. Given that an entry may contain more than one retrieved predicate for its noun, we scaled the dictionary to allow for duplicate nouns with different predicates (up to three predicates).[3] This yielded over 20,000 entries.

Sentence construction: For each of these adjective-noun-predicate entries, we generate natural sentences and four corresponding variants. The original sentence (s in Section 3) is composed only from the noun and the predicate, while the four variants (s' in Section 3) are modified versions of the original sentence created by adding the adjective before the root noun in the original sentence

[2] Preliminary analyses showed that non-subsective adjectives represent less than 5% of our entries.

[3] Threshold selected to correspond to the average number of different predicates extracted for each noun, and avoid scale the dictionary excessively at the cost of dataset diversity.

Noun-amod Extraction:	**Amod**: *victorious* **Noun**: *candidate* **Count**: 181	**Amod**: *third* **Noun**: *candidate* **Count**: *222*	**Amod**: *qualified* **Noun**: *candidate* **Count**: *250*	**Amod**: *false* **Noun**: *candidate* **Count**: *130*
Predicate Extraction:	**Noun**: *candidate* **Predicate**: *win an election (Relation: IsCapableOf)* **Surface Text**: *[[A candidate]] can [[win an election]]*			
Sentence Construction:	**Original Sentence**: *A candidate wins an election.* **Variant 1**: *A [victorious] candidate wins an election.* **Variant 2**: *A [third] candidate wins an election.* **Variant 3**: *A [qualified] candidate wins an election.* **Variant 4**: *A [false] candidate wins an election.*			
Label Generation:	**Original Sentence**: *A candidate wins an election.* **Modified Sentence**: *A [victorious] candidate wins an election.*			**Label**: *Necessarily true*

Table 2: Examples of how ADEPT instances are created through the pipeline.

Prompt:	**Compared with the original statement** (*"A clock is working correctly."*) **please assess the plausibility of the following modified version:** *"A **broken** clock is working correctly."*				
Plausibility change:	✓ Impossible	✗ Less Likely	✗ Equally Likely	✗ More Likely	✗ Necessarily True
Prompt:	**Compared with the original statement** (*"A candidate wins the election."*) **please assess the plausibility of the following modified version:** *"A **third** candidate wins the election."*				
Plausibility change:	✗ Impossible	✗ Less Likely	✓ Equally Likely	✗ More Likely	✗ Necessarily True
Prompt:	**Compared with the original statement** (*"A mistake angers a person."*) **please assess the plausibility of the following modified version:** *"A **fatal** mistake angers a person."*				
Plausibility change:	✗ Impossible	✗ Less Likely	✗ Equally Likely	✓ More Likely	✗ Necessarily True

Table 3: Examples of how ADEPT instances are labelled in the crowdsourcing interface.

(see Table 2 for examples). To create the natural sentences themselves, we modify the surface text by replacing modal verbs (like *can* or *may*) with the declarative *is*, as modal verbs may complicate the evaluation of what the plausibility of described events might be.

Adjective Sampling: To identify non-subsective adjectives in the dataset entries, we use a set of 60 non-subsective adjectives identified by Nayak et al. (2014). Then, to select four adjectives we first 1) randomly select up to two non-subsective modifiers if they co-occured with the noun, and then 2) we randomly select the remaining adjectives from the list of subsective modifiers. We over-sample non-subsective modifiers as they occur sparsely in the corpora and we want to evaluate their effects against other modifiers. This random sampling strategy results in an about 1:4 non-subsective to subsective adjective ratio (as some entries have no non-subsective adjectives), allowing us to analyze the effect of non-subsective modifiers while maintaining an element of randomness.

Label Generation + Quality Control: For each entry, annotators (from Mechanical Turk) label one randomly selected sentence variant (from the four variants) against its original sentence, with labels indicating the change in plausibility due to adding the selected adjective (Table 3). For quality control, we also add roughly 2,000 quality-check entries—including gold label instances for which there was unanimous agreement among four annotators in earlier pilots and "attention-check" instances that explicitly ask annotators to select a specific label. We filter out all instances annotated by annotators who failed the attention checks or whose labels differed by at least two degrees from the gold labels (e.g., selected *equally likely* when the gold label was *impossible*) on more than 10% of their annotations. We also limit the maximum number of labelling tasks per annotator to 100 (corresponding to less than 0.5% of the data) to ensure that no one judge significantly affects the quality of the data. Finally, we only keep those instances for which we observe a majority agreement (i.e., at least two annotators agree about the final label). After this final quality-control filtering steps, the final dataset includes 16,115 instances.

Agreement	Impossible	Less Likely	Eq. Likely	More Likely	Nec. True
Unanimous (5267)	0.21	0.16	0.40	0.15	0.09
Majority (10848)	0.79	0.84	0.60	0.85	0.91
No Agreement (3209)	–	–	–	–	–
Label Distribution	0.14	0.12	0.67	0.07	0.01
Dataset Split (16115)		Train Set: 12892	Val Set: 1611	Test Set: 1612	

Table 4: Dataset statistics in terms of agreement, label, and size distributions. The stats for Unanimous & Majority represent their prevalence among the pairs for which we observed a majority agreement (and sum up to 1).

4.2 Dataset Quality Assessment

Table 4 overviews the dataset figures, highlighting the labels' distribution and agreement. By inspecting how often judges agree across our plausibility labels, we observe higher assessment variability for instances with labels further from *equally likely* (also the most commonly applied label). This is particularly true for instances with labels at the extremes of our plausibility scale (i.e., the *impossible* and *necessarily true* labels). While 40% of the dataset instances marked as *equally likely* have unanimous annotator agreement, this is the case for only 21% of the instances marked as *impossible*. We found no agreement across the 5 plausibility labels (§3) for about 15% of the annotated instances, which we do not include in the final dataset.

While how much judges agree on labels varies across plausibility levels, the directionality of the assigned labels is more stable—i.e., many disagreements are due to judges making different but consistent assessments like *more likely* and *necessarily true*, rather than conflicting assessments like *less* and *more likely*. Because of this, we also experiment with alternative 3-Class and 4-Class task formulations (§5.3), where the *impossible* and *necessarily true* labels are either combined with other labels or are discarded.

Task Ambiguity Closely inspecting instances marked as *impossible* and *necessarily true* to understand possible sources of disagreement among judges, we find that only about a quarter (for *impossible*) to a third (for *necessarily true*) of these instances appear to be clear cases where both **1)** adding the adjective led to a change in plausibility **and 2)** the change in plausibility made the event *impossible* or *necessarily true*.

Sometimes the described events are already *impossible* or *necessarily true* (e.g., *average* in "an [average] week is made up of seven days" does not change the plausibility of this statement, which was already *necessarily true*). In other cases, the added modifier changes the semantic interpretation of the event (e.g., the modifier *algebraic* makes the event "an [algebraic] operator pages a doctor" impossible because it alters the sense of the term *operator*), or it introduces grammatical or logical errors (e.g., "[former] sleeping is for maintaining sanity" was likely marked as impossible for being illogical). There are also clear cases of false positives, where the resulting events are not *impossible* or *necessarily true* (e.g., "[romantic] Jasmine buys her dress at the store" is not *impossible*).

These issues were particularly prevalent among instances annotated as *impossible*, where about half of the instances appear to be false positives, ungrammatical, or nonsensical sentences. We therefore also experiment with a 4-Class formulation that does not include the *impossible* label (§5.3).

Task Reliability Given the subjective and ambiguous nature of our task, we also sought to characterize to what extent the overall reliability of our labels might be affected by it. For this, two authors independently labelled 100 randomly sampled instances from ADEPT, using the same annotation specifications provided to the crowdsourcing judges. We then measured the inter-assessor agreement between the two authors Cohen's Kappa $\kappa = 0.82$, which indicates substantial agreement.

We then take the instances where both authors agreed (87%) and compare their labels with those provided by the crowd-workers, obtaining a $\kappa = 0.74$ that while lower is still substantial. Finally, the individual agreement of each of the authors' labels with crowdsourcing judges (which includes cases where authors disagree) corresponded to $\kappa = 0.77$ and $\kappa = 0.64$, further demonstrating the *overall reliability* of the labels we collected.

Model	3-Class Dev. Accuracy	5-Class Dev. Accuracy
Majority Prediction	66.4	66.4
Normative Rule	70.1	63.6
Human	90.0[4]	85.0[4]
Human (no context)	75.0[4]	71.0[4]
BERT (no context)	72.0	69.4
RoBERTa (no context)	72.4	69.1
DeBERTa (no context)	72.1	68.6
BERT	72.3	69.8
RoBERTa	73.1	**70.8**
DeBERTa	**73.9**	69.7

Table 5: Performance of various models on the ADEPT development set.

5 Methods

5.1 Neural Models

We evaluate several transformer-based models on ADEPT. For fine-tuning, we adopt the standard practice for sentence-pair tasks described by Devlin et al. (2015). We concatenate the first and second sentence with *[SEP]*, prepend the sequence with *[CLS]*, and feed the input to the transformer model. The representation for *[CLS]* is fed into a softmax layer for a five-way classification.

BERT (Devlin et al., 2015) is one of the first transformer-based architectures, featuring a pre-trained neural language model with bidirectional paths and sentence representations in consecutive hidden layers.

RoBERTa (Liu et al., 2019) is an improved variant of BERT that adds more training data with larger batch sizes and longer training, as well as other refinements like dynamic masking. RoBERTa performs consistently better than BERT across many benchmarks (Wang et al., 2018a).

DeBERTa builds on RoBERTa with disentangled attention and enhanced mask decoder training with half the data used in RoBERTa; currently the best-performing transformer-based model on several NLI-related tasks (He et al., 2020).

5.2 Baseline Models

Majority Prediction This heuristic always predicts the *equally likely* label, which represents 67% of the full dataset.

Normative Rule This heuristic corresponds to the normative treatment of non-subsective modifiers according to the taxonomy described in Section 2, where the general expectation is that the insertion of a non-subsective adjective would reduce the plausibility of the modified sentence. Thus, when the inserted adjective in s' is among the list of non-subsective modifiers, this baseline predicts *less likely*, otherwise it predicts the majority label, which is *equally likely*.

No Context Baseline We run a word association baseline to evaluate to what extent context is needed to solve the dataset. In this baseline, the transformer model is provided only the noun from s' as the representation for the original sentence, and the modifier a as the representation for the the modified sentence s separated by *[SEP]* (e.g., for sentence 1a in the introduction, this corresponds to the input: *monkey [SEP] dead*). This is analogous to the hypothesis-only baseline in NLI (Belinkov et al., 2019), where the task does not require the full context to achieve high performance.

Human Evaluation To estimate human performance on our task, a new annotator (not an author) independently assessed a random sample of 100 validation instances from ADEPT. The annotator then evaluated each sentence using the same instructions provided to the crowdsourcing judges, whose majority agreement determined the final label. The human performance thus corresponds to the percentage of instances for which the new annotator's labels agree with the ADEPT labels. We also estimate human performance under a *no context* setting, where we presented this same annotator (who was now well-acquainted with the task) with a new random sample with only the noun and the modifier. The annotator then made their best guess as to what the plausibility difference was without knowing the context. We ensured the new instances were distinct from those in the first random sample.

5.3 Experiments

We primarily test the baselines and transformer models using two metrics. The first metric corresponds to the prediction accuracy on the full five-label classification task (5-Class Accuracy). As an alternative metric—drawing from our observations in Section 4.2—we use the accuracy on a three-label classification task (3-Class Accuracy), where we bundle *impossible* and *less likely* into a single label representing *a decrease in plausibility*, and *necessarily true* and *more likely* into a label representing an *increase in plausibility*.

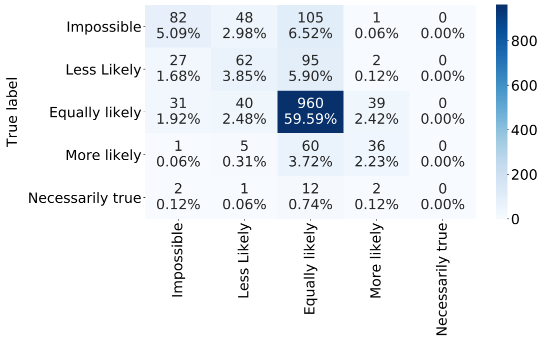

Figure 2: Confusion matrix for the best model on the 5-class setting (RoBERTa), ADEPT development set.

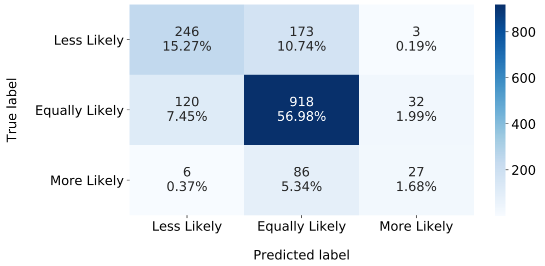

Figure 3: Confusion matrix for the best model on the 3-class setting (DeBERTa), ADEPT development set.

5.4 Training

All models are implemented and trained using HuggingFace's Transformers library (Wolf et al., 2020). We use grid-search for hyper-parameter tuning: learning rate {1e-5, 3e-5, 5e-5}, number of epochs {3, 4, 5, 8}, batch-size {8, 16, 32} with three different random seeds. For fine-tuning, we allow for the fine-tuning of all parameters including those in the model's hidden layers.[3]

6 Results

Easy for Humans, Difficult for Transformers: Model prediction accuracy is summarized in Table 5, where the general trend is as follows: the transformer-based models have a higher prediction accuracy than the majority prediction and normative rule baselines, but still fall short of human performance by a large margin.

Of the transformer models, the highest 3-class accuracy is achieved by DeBERTa and the highest 5-class accuracy by RoBERTa; however, the difference in accuracy of all transformer models is small (and not statistically significant p-value > 0.05), being within a 1.6% range.

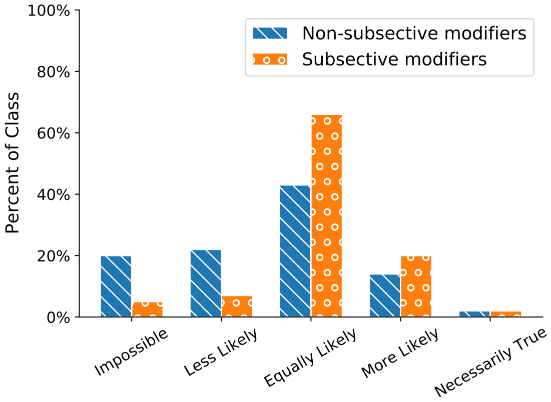

Figure 4: Label distribution of ADEPT according to taxonomic class of modifier.

In the *no-context* ablations where models only see the noun phrase and modifier, the transformer models performance decreases only slightly, which suggests the models might be "insensitive" to context. In contrast, approximated human performance decreases significantly in the no-context setting, dropping e.g., from 90% to 75% accuracy for 3-class predictions. This no-context human accuracy, however, is still superior to the best performing transformer model with context.

To understand what errors the models make, we examine the confusion matrices for the best performing models on both the 3-class (Figure 2) and 5-class formulations (Figure 3). The most common errors appear to happen when a change in plausibility is erroneously classified as *equally likely*, and when a modifier that does not change an event's plausibility is erroneously predicted to render the new sentence as *less likely*. Table 6 includes example sentences along with 5-Class predictions by the best performing transformer model.

The Taxonomic Classes Just Don't Cut it: Figure 4 shows the distribution of plausibility labels in ADEPT for both subsective and non-subsective modifiers. We see that both classes of modifiers lead to a wide mix of changes in the plausibility of given events, corroborating Pavlick and Callison-Burch (2016)'s findings that normative rules cannot categorically describe a modifier's behavior. This likely also explains the poor performance of the *normative rule* baseline on both the 5- or 3-class plausibility classification task formulations.

Ambiguity Makes Operationalizing Plausibility Difficult: Some of our plausibility labels prove

[3] We also evaluated models where we froze the parameters of all the hidden layers as a probing mechanism, but found that no model performed better than the majority baseline.

[4] This is an estimate based on a subsample of the data.

ADEPT instance	Annotated Change	RoBERTa Prediction	Correct?
A [professional] mathematician proves a theorem.	More Likely	More Likely	✓
[Questionable] evidence proves innocence.	Less Likely	Less Likely	✓
A [western] kitchen is for storing food.	Equally Likely	Equally Likely	✓
[Yellow] bananas are yellow.	Necessarily True	Equally Likely	✗
You use a [strong] jack to lift your car.	More Likely	Equally Likely	✗
A [dead] leg has one foot.	Equally Likely	Impossible	✗

Table 6: Representative examples from the development set and corresponding model predictions.

ambiguous and harder to reliably assign to our dataset instances, particularly at the extremes of our plausibility scale (§4.2). Given that for the *impossible* label many instances did not appear to correctly capture changes in plausibility that render the modified event *impossible*, we conduct exploratory experiments with a 4-Class task formulation that excludes the impossible class. For the best performing model (RoBERTa), we observe an overall improved accuracy from 70.8% to 81.2% (compared to the 5-Class classification task).

Better plausibility classification schemes and crowdsourcing protocols might help us more effectively operationalize plausibility changes. However, how to effectively separate between 1) cases where the modifiers alter the semantic interpretation of a statement (and thus lead to a different event) or make the sentences ungrammatical *versus* 2) cases where modifiers actually lead to changes in the plausibility of the original event, remains an open question.

7 Conclusions

We present a new large-scale corpus and task, ADEPT, for assessing semantic plausibility. Our corpus contains over 16 thousand difficult task instances, specifically constructed to test a system's ability to correctly interpret and reason about adjective-noun pairs within a given context. Our experiments suggest a persistent performance gap between human annotators and large language representation models, with the later exhibiting a lower sensitivity to context. Finally, our task provides deeper insight into the effects of various classes of adjectives on event plausibility, and suggests that rules based solely on the adjective or its denotation do not suffice in determining the correct plausibility readings of events.

In the future, we wish to investigate how ADEPT could be used to improve performance on related natural language inference tasks (e.g. MNLI, SNLI & SciTail (Khot et al., 2017)). We also plan to develop new models on ADEPT and transfer them to other semantic plausibility tasks.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and by Microsoft Research. Jackie Chi Kit Cheung is supported by the Canada CIFAR AI Chair program, and is also a consulting researcher for Microsoft Research.

Ethical Considerations

While our focus on examining what effects adjectives have on the plausibility of arbitrary events makes ascertaining the broader impact of our work challenging, this work is *not* void of possible adverse social impacts or unintended consequences.

First, to generate our dataset of events, we use English Wikipedia, Common Crawl, and ConceptNet5 (based on data from e.g., Games with a Purpose or DBPedia). Such data sources are however known to exhibit a range of biases (Olteanu et al., 2019; Baeza-Yates, 2018)—which LMs reproduce (Solaiman et al., 2019)—being often unclear what and whose content they represent. While our goal is to enable others to explore the effects of modifiers and how these effect might impact various inference tasks, users of this dataset should acknowledge possible biases and should not use it to make deployment decisions or rule out failures. To this end, our dataset release will be accompanied by a datasheet (Gebru et al., 2018).

Depending on the context, determining changes in plausibility can also be ambiguous or even subjective (see §4.2). This means that in some downstream applications, possible plausibility inference errors might, for instance, inadvertently elevate factually incorrect, subjective or misleading beliefs. If those inference errors happen more when events concern certain groups or activities, they might have disparate effects across stakeholders. Thus, understanding the potential impact of our plausibility inference task requires us to think about both downstream applications and possible stakeholders (Boyarskaya et al., 2020). For instance, one application of plausibility inferences is perhaps veracity or credibility assessment. It would be problematic if a system would reproduce highly harmful stereotypes by inferring that a <u>*black*</u> *witness* is *less likely* to be trustworthy than just a *witness*, or that an <u>*old*</u> *applicant* is *less likely* to be a productive employee than just an *applicant*. Another application (we also used as a motivating example) is information extraction where perhaps such plausibility inferences could be used to infer which details to keep during extraction. Errors might for instance harmfully reinforce the belief that the prototypical human is male (Menegatti and Rubini, 2017), if <u>*female*</u> is deemed as more likely to change the plausibility of events about e.g., doctors, scientists, or other professionals; and thus deemed a relevant (or not) detail to surface based on it.

References

Marilisa Amoia and Claire Gardent. 2006. Adjective based inference. In *Proceedings of the Workshop KRAQ'06: Knowledge and Reasoning for Language Processing*.

Marilisa Amoia and Claire Gardent. 2007. A first order semantic approach to adjectival inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 185–192.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming failures of imagination in AI infused system development and deployment. In *Navigating the Broader Impacts of AI Research Workshop at NeurIPS 2020*.

Romane Clark. 1970. Concerning the logic of predicate modifiers. *Nous*, pages 311–335.

Ido Dagan, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct word sense matching for lexical substitution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 449–456, Sydney, Australia. Association for Computational Linguistics.

Guillermo Del Pinal. 2015. Dual content semantics, privative adjectives, and dynamic compositionality. *Semantics and Pragmatics*, 8:7–1.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. 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 100–105, Beijing, China. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Wenguan Huang and Xudong Luo. 2017. Commonsense reasoning in a deeper way: By discovering relations between predicates. In *ICAART*.

Hans Kamp and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57(2):129–191.

J. A. W. Kamp and Edward L. Keenan. 1975. *Two theories about adjectives*, page 123–155. Cambridge University Press.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 311–316, Vancouver, Canada. Association for Computational Linguistics.

Maria Lapata, Scott McDonald, and Frank Keller. 1999. Determinants of adjective-noun plausibility. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, Bergen, Norway. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

John Philip McCrae, Francesca Quattri, Christina Unger, and Philipp Cimiano. 2014. Modelling the semantics of adjectives in the ontology-lexicon interface. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 198–209.

Louise McNally and Gemma Boleda. 2004. Relational adjectives as properties of kinds. *EMPIRICAL*, page 179.

Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*.

Richard Montague. 1970. English as a formal language. In Bruno Visentini, editor, *Linguaggi nella societa e nella tecnica*, pages 188–221. Edizioni di Communita.

James Mullenbach, Jonathan Gordon, Nanyun Peng, and Jonathan May. 2019. Do nuclear submarines have nuclear captains? a challenge dataset for commonsense reasoning over adjectives and objects. 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 6052–6058, Hong Kong, China. Association for Computational Linguistics.

Neha Nayak, Mark Kowarsky, Gabor Angeli, and Christopher D Manning. 2014. A dictionary of nonsubsective adjectives. Technical report, Technical Report CSTR 2014-04, Department of Computer Science, Stanford.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone P. Ponzetto, and Chris Biemann. 2018. Building a web-scale dependency-parsed corpus from CommonCrawl. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Terence Parsons. 1970. Some problems concerning the logic of grammatical modifiers. *Synthese*, 21(3):320–334.

Ellie Pavlick and Chris Callison-Burch. 2016. So-called non-subsective adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 114–119.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Roman Inflianskas. 2020. profanity-filter.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

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

Su Wang, Greg Durrett, and Katrin Erk. 2018b. Modeling semantic plausibility by injecting world knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 303–308, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.