Variation and generality in encoding of syntactic anomaly information in sentence embeddings

Qinxuan Wu[*]
College of Computer Science and Technology
Zhejiang University
wuqinxuan@zju.edu.cn

Allyson Ettinger
Department of Linguistics
The University of Chicago
aettinger@uchicago.edu

Abstract

While sentence anomalies have been applied periodically for testing in NLP, we have yet to establish a picture of the precise status of anomaly information in representations from NLP models. In this paper we aim to fill two primary gaps, focusing on the domain of syntactic anomalies. First, we explore fine-grained differences in anomaly encoding by designing probing tasks that vary the hierarchical level at which anomalies occur in a sentence. Second, we test not only models' ability to detect a given anomaly, but also the generality of the detected anomaly signal, by examining transfer between distinct anomaly types. Results suggest that all models encode some information supporting anomaly detection, but detection performance varies between anomalies, and only representations from more recent transformer models show signs of generalized knowledge of anomalies. Follow-up analyses support the notion that these models pick up on a legitimate, general notion of sentence oddity, while coarser-grained word position information is likely also a contributor to the observed anomaly detection.

1 Introduction

As the NLP community works to understand what is being learned and represented by current models, a notion that has made sporadic appearances is that of linguistic anomaly. Analyses of language models have often tested whether models prefer grammatical over ungrammatical completions (e.g. Linzen et al., 2016), while analyses of sentence embeddings have probed for syntax and semantics by testing detection of sentence perturbations (Conneau et al., 2018). Such work tends to exploit anomaly detection as a means of studying linguistic phenomena, setting aside any direct questions about encoding of anomaly per se. However, models' treatment of anomaly is itself a topic that raises

important questions. After all, it is not obvious that we should expect models to encode information like "this sentence contains an anomaly", nor is it obvious which types of anomalies we might expect models to pick up on more or less easily. Nonetheless, anomalies are easy to detect for humans, and their detection is relevant for applications such as automatic error correction (Ge et al., 2018), so it is of value to understand how anomalies operate in our models, and what impacts anomaly encoding.

In the present work we seek to fill this gap with a direct examination of anomaly encoding in sentence embeddings. We begin with fine-grained testing of the impact of anomaly type, designing probing tasks with anomalies at different levels of syntactic hierarchy to examine whether model representations better support detection of certain types of anomaly. Then we examine the generality of anomaly encoding by testing transfer performance between distinct anomalies—here our question is, to the extent that we see successful anomaly detection, does this reflect encoding of a more general signal indicating "this sentence contains an anomaly", or does it reflect encoding of simpler cues specific to a given anomaly? We focus on syntactic anomalies because the hierarchy of sentence structure is conducive to our fine-grained anomaly variation. (Sensitivity to syntactic anomalies has also been studied extensively as part of the human language capacity (Chomsky, 1957; Fodor et al., 1996), strengthening precedent for prioritizing it.)

We apply these tests to six prominent sentence encoders. We find that most models support non-trivial anomaly detection, though there is substantial variation between encoders. We also observe differences between hierarchical classes of anomaly for some encoders. When we test for transferability of the anomaly signal, we find that for most encoders the observed anomaly detection shows little sign of generality—however, transfer performance in BERT and RoBERTa suggests

---

[*]This work was done while the first author was a visiting student at the University of Chicago.

that these more recent models may in fact pick up on a generalized awareness of syntactic anomalies. Follow-up analyses support the possibility that these transformer-based models pick up on a legitimate, general notion of syntactic oddity—which appears to coexist with coarser-grained, anomaly-specific word order cues that also contribute to detection performance. We make all data and code available for further testing.[1]

2 Related Work

This paper builds on work analyzing linguistic knowledge reflected in representations and outputs of NLP models (Tenney et al., 2019; Rogers et al., 2020; Jawahar et al., 2019). Some work uses tailored challenge sets associated with downstream tasks to test linguistic knowledge and robustness (Dasgupta et al., 2018; Poliak et al., 2018a,b; White et al., 2017; Belinkov et al., 2017b; Yang et al., 2015; Rajpurkar et al., 2016; Jia and Liang, 2017; Rajpurkar et al., 2018). Other work has used targeted classification-based probing to examine encoding of specific types of linguistic information in sentence embeddings more directly (Adi et al., 2016; Conneau et al., 2018; Belinkov et al., 2017a; Ettinger et al., 2016, 2018; Tenney et al., 2019; Klafka and Ettinger, 2020). We expand on this work by designing analyses to shed light on encoding of syntactic anomaly information in sentence embeddings.

A growing body of work has examined syntactic sensitivity in language model outputs (Chowdhury and Zamparelli, 2018; Futrell et al., 2019; Lakretz et al., 2019; Marvin and Linzen, 2018; Ettinger, 2020), and our *Agree-Shift* task takes inspiration from the popular number agreement task for language models (Linzen et al., 2016; Gulordava et al., 2018; Goldberg, 2019). Like this work, we focus on syntax in designing our tests, but we differ from this work in focusing on model representations rather than outputs, and in our specific focus on understanding how models encode information about anomalies. Furthermore, as we detail below, our *Agree-Shift* task differs importantly from the LM number agreement tests, and should not be compared directly to results from those tests.

Our work relates most closely to studies involving anomalous or erroneous sentence information (Warstadt et al., 2019; Yin et al., 2020; Hashemi

and Hwa, 2016). Some work investigates impacts from random shuffling or other types of distortion of input text (Pham et al., 2020; Gupta et al., 2021) or of model pre-training text (Sinha et al., 2021) on downstream tasks—but this work does not investigate models' encoding of these anomalies. Warstadt et al. (2019) present and test with the *CoLA* dataset for general acceptability detection, and among the probing tasks of Conneau et al. (2018) there are three that involve analyzing whether sentence embeddings can distinguish erroneous modification to sentence inputs: *SOMO*, *BShift*, and *CoordInv*. Yin et al. (2020) also generate synthetic errors based on errors from non-native speakers, showing impacts of such errors on downstream tasks, and briefly probing error sensitivity. More recently, Li et al. (2021) conduct anomaly detection with various anomaly types at different layers of transformer models, using training of Gaussian models for density estimation, and finding different types of anomaly sensitivity at different layers. We build on this line of work in anomaly detection with a fine-grained exploration of models' detection of word-content-controlled perturbations at different levels of syntactic hierarchy. Our work is complementary also in exploring generality of models' anomaly encoding by examining transfer performance between anomalies.

3 Syntactic Anomaly Probing Tasks

To test the effects of hierarchical location of a syntactic anomaly, we create a set of tasks based on four different levels of sentence perturbation. We structure all perturbations so as to keep word content constant between original and perturbed sentences, thus removing any potential aid from purely lexical contrast cues. Our first three tasks involve reordering of syntactic constituents, and differ in hierarchical proximity of the reordered constituents: the first switches constituents of a noun phrase, the second switches constituents of a verb phrase, and the third switches constituents that only share the clause. Our fourth task tests sensitivity to perturbation of morphological number agreement, echoing existing work testing agreement in language models (Linzen et al., 2016).

3.1 *Mod-Noun*: Detecting modifier/noun reordering

Our first task tests sensitivity to anomalies in modifier-noun structure, generating anomalous sen-

tences by swapping the positions of nouns and their accompanying modifiers, as below:

*A man wearing a **yellow scarf** rides a bike.* →
*A man wearing a **scarf yellow** rides a bike.*

We call this perturbation *Mod-Noun*. Any article determiner of the noun phrase remains unperturbed.

3.2 *Verb-Ob*: Detecting verb/object reordering

Our second task tests sensitivity to anomalies in English subject-verb-object (SVO) sentence structure by swapping the positions of verbs and their objects (SVO → SOV). To generate perturbed sentences for this task, we take sentences with a subject-verb-object construction, and reorder the verb (or verb phrase) and the object, as in the example below:

*A man wearing a yellow scarf **rides a bike**.* →
*A man wearing a yellow scarf **a bike rides**.*

We refer to this perturbation as *Verb-Ob*. Note that *Verb-Ob* and *Mod-Noun* are superficially similar tasks in that they both reorder sequentially consecutive constituents. However, importantly, they differ in the hierarchical level of the swap.

3.3 *SubN-ObN*: Detecting subject/object reordering

Our third task tests sensitivity to anomalies in subject-verb-object relationships, creating perturbations by swapping the positions of subject and object nouns in a sentence. For this task, we generate the data by swapping the two head nouns of the subject and the object, as below:

*A **man** wearing a yellow scarf rides a **bike**.* →
*A **bike** wearing a yellow scarf rides a **man**.*

We refer to this perturbation as *SubN-ObN*. We target subject-verb-object structures directly under the root of the syntactic parse, meaning that only one modification is made per sentence for this task.

Detecting the anomaly in this perturbation involves sensitivity to argument structure (the way in which subject, verb, and object should be combined), along with an element of world knowledge (knowing that a bike would not ride a man, nor would a bike typically wear a scarf).[2]

---

[2]For more details about this task, please see Appendix A.7.

3.4 *Agree-Shift*: Detecting subject/verb disagreement

Our fourth task tests sensitivity to anomalies in subject-verb morphological agreement, by changing inflection on a present tense verb to create number disagreement between subject and verb:

*A **man** wearing a yellow scarf **rides** a bike.* →
*A **man** wearing a yellow scarf **ride** a bike.*

We refer to this perturbation as *Agree-Shift*.[3] This is the only one of our tasks that involves a slight change in the word inflection, but the word stem remains the same—we consider this to be consistent with holding word content constant.

4 Experiments

We generate probing datasets for each of the anomaly tests described above. We then apply these tasks to examine anomaly sensitivity in a number of generic sentence encoders.[4]

Datasets Each of the above perturbations is used to create a probing dataset consisting of normal sentences and corresponding modified sentences, labeled as *normal* and *perturbed*, respectively. Within each probing task, each *normal* sentence has a corresponding *perturbed* sentence, so the label sets for each task are fully balanced. Each probe is formulated as a binary classification task. Normal sentences and their corresponding perturbed sentences are included in the same partition of the train/dev/test split, so for any sentence in the test set, no version of that sentence (neither the perturbed form nor its original form) has been seen at training time. We draw our *normal* sentences from MultiNLI (Williams et al., 2018) (premise only). Perturbations of those sentences are then generated as our *perturbed* sentences.

---

[3]Note that while this perturbation echoes the popular LM agreement analyses (Linzen et al., 2016; Gulordava et al., 2018; Goldberg, 2019), the fact that we are probing sentence embeddings for explicit detection of this anomaly is an important difference. Performance by LMs on those agreement tasks can indicate that a model prefers a non-anomalous completion, but cannot speak to whether the model encodes any explicit/perceptible awareness that an anomaly is present/absent. For this reason, model performance on our *Agree-Shift* task should not be compared directly to performance on these agreement probability tasks.

[4]Please refer to the Appendix A.5, A.6, A.7 for more details about data generation, probing implementation, as well as descriptions about encoders and external tasks.
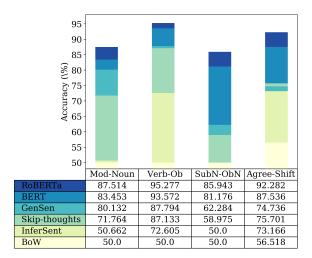
	Mod-Noun	Verb-Ob	SubN-ObN	Agree-Shift
RoBERTa	87.514	95.277	85.943	92.282
BERT	83.453	93.572	81.176	87.536
GenSen	80.132	87.794	62.284	74.736
Skip-thoughts	71.764	87.133	58.975	75.701
InferSent	50.662	72.605	50.0	73.166
BoW	50.0	50.0	50.0	56.518

Figure 1: Anomaly detection performance.

Probing We analyze sentence embeddings from these prominent sentence encoders: InferSent (Conneau et al., 2017), Skip-thoughts (Kiros et al., 2015), GenSen (Subramanian et al., 2018), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019b). The first three models are RNN-based, while the final two are transformer-based.

To test the effectiveness of our control of word content, we also test bag-of-words (BoW) sentence embeddings obtained by averaging of GloVe (Pennington et al., 2014) embeddings. This allows us to verify that our probing tasks are not solvable by simple lexical cues (c.f. Ettinger et al., 2018), thus better isolating effects of syntactic anomalies.

We train and test classifiers on our probing tasks, with sentence embeddings from the above encoders as input. The classifier structure is a multilayer perceptron (MLP) classifier with one hidden layer.[5]

5 Anomaly Detection Results

Fig. 1 shows anomaly detection performance for the tested encoders. We can see first that for three of our four tasks—all reordering tasks—our BoW baseline performs perfectly at chance, verifying elimination of lexical biases. BoW on the *Agree-Shift* task is just above chance, reflecting (expected) slight bias in morphological variations.

Comparing between tasks, we see that *Verb-Ob* yields highest overall performance while *SubN-ObN* yields the lowest. As mentioned above, a particularly informative comparison is between *Verb-Ob* and *Mod-Noun*, which both involve swapping sequentially adjacent content, but at different hier-

archical levels. We see that encoders consistently show stronger performance on *Verb-Ob* than *Mod-Noun*, suggesting that the broader hierarchical domain of *Verb-Ob* may indeed make anomalies more accessible for encoding. The only anomaly that affects a broader span of the sentence is *SubN-ObN*— but we see that this is instead one of the most challenging tasks. We suspect that this is attributable to the fact that, as described above, detecting this anomaly may require extra world knowledge and common sense, which certain encoders may have less access to.[6] It is not unexpected, then, that BERT and RoBERTa, with comparatively much larger and more diverse training data exposure, show a large margin of advantage on this challenging *SubN-ObN* task relative to the other encoders. *Agree-Shift* patterns roughly on par with *Mod-Noun*, though notably InferSent (and Skip-thoughts) detects the agreement anomaly much more readily than it does the *Mod-Noun* anomaly.

Comparing between encoders, we see clear stratification in performance. InferSent shows the least anomaly awareness, performing for half of the tasks at chance level with the BoW baseline. GenSen and Skip-thoughts, by contrast, consistently occupy a higher performance tier, often falling not far behind (but never quite on par with) the highest level of performance. The latter distinction is reserved for BERT and RoBERTa, which show the strongest anomaly sensitivity on all tasks. All models show stronger performance on *Verb-Ob* than *Mod-Noun*, but the hierarchical difference between these tasks seems to have particularly significant impact for InferSent and Skip-thoughts, with cues relating to *Verb-Ob* seemingly encoded by InferSent, but cues relating to *Mod-Noun* seeming to be absent. *Mod-Noun* also yields the largest margin of difference between GenSen and Skip-thoughts. Since Skip-thoughts is one objective of GenSen, this suggests that the additional GenSen objectives provide an edge particularly for the finer-grained information needed for *Mod-Noun*.

BERT and RoBERTa emerge soundly as the strongest encoders of anomaly information, with RoBERTa also consistently outperforming BERT. While this is in line with patterns of downstream task performance from these models, it is noteworthy that these models also show superior perfor-

---

[5]We also train on a logistic regression (LR) classifier. LR results are shown in the Appendix Table 4.

[6]As an encoder trained on semantic reasoning, InferSent nonetheless fails terribly on this task—this may be explained by findings that heuristics can account for much of NLI task learning (Poliak et al., 2018c; McCoy et al., 2020).

mance on these anomaly-detection tasks, as it is not obvious that encoding anomaly information would be relevant for these models' pre-training objectives, or for the most common NLU tasks on which they are typically evaluated.

6 Investigation on Generality of Anomaly Encoding

The above experiments suggest that embeddings from many of these encoders contain signal enabling detection of the presence of syntactic anomalies. However, these results cannot tell us whether these embeddings encode awareness that an "error" is present per se—the classifiers may simply have learned to detect properties associated with a given anomaly, e.g., agreement mismatches, or occurrence of modifiers after nouns. In this sense, the above experiments serve as finer-grained tests of levels of hierarchical information available in these embeddings, but still do not test awareness of the notion of anomaly in general.

In this section we take a closer look at anomaly awareness per se, by testing the extent to which the sensitivities identified above are specific to individual anomalies, or reflective of a more abstract "error" signal that would apply across anomalies. We explore this question by testing transfer performance between different anomaly types.

Transfer Results   While in Section 5 we focused on examining anomaly-specific sensitivity in our new tasks—testing variation along fine-grained syntactic hierarchical distinctions and in a word-controlled setting—for examining generality of anomaly encoding it is worthwhile to take into account a broader range of anomaly types and datasets. For this reason we examine transfer between each of our generated probing tasks, as well as transfer to our tasks from established datasets: *SOMO*, *BShift*, and *CoordInv* from Conneau et al. (2018), and the *CoLA* task (Warstadt et al., 2019).

Fig. 2 shows transfer results from each dataset to each of our tasks. For ease of comparison, we also show the test result achieved when training on the same anomaly (the non-transfer result) in the white bars. We see that the majority of encoders show a marked drop in performance relative to the original non-transfer accuracy, and in fact most of the transfer results are approximately at chance performance. This suggests that the embeddings from these models encode information supporting detection of these anomalies, but the signals that

enable this detection are anomaly-specific. That is, they may encode some syntactic/semantic signal supporting detection of specific anomalies, but there is no indication that they encode a general awareness that "there is an anomaly".

The notable exceptions to this poor transfer performance are the transformer-based models, BERT and RoBERTa, which by stark contrast to the RNN-based encoders, show non-trivial transfer performance across all four of our generated tasks, regardless of the anomaly that the classifier is trained on. This suggests that to a much greater extent than any of the other encoders, BERT and RoBERTa may encode a more general "error" signal, allowing for generalization across anomalies. Importantly, BERT and RoBERTa do also show some performance drop from non-transfer to transfer settings—so while they may encode a more generalized "error" signal, this is likely in combination with encoding of anomaly-specific information that further aids performance on a given anomaly.

We note that transfer performance from *CoLA* is typically comparable to, or occasionally better than, training on the Conneau et al. tasks–despite the fact that *CoLA* has much smaller training data. *CoLA* is also the only task that contains a variety of anomalies for the model to learn from, rather than a single anomaly type as in all other datasets. This may enable faster, more generalizable learning on this small dataset—but of course, this would only be possible if a generalized anomaly signal is available in the embeddings. Following this reasoning, we also test whether jointly training on multiple anomalies improves transfer performance. The results of these multi-task transfer experiments can be found in Appendix Tables 7-8. These transfer results show overall a small decrease in performance relative to the one-to-one transfer, suggesting that training on single types of anomalies is not creating any major disadvantage for transfer performance. It may also indicate that mixed types of higher-level oddity in natural occurring anomalies from *CoLA* is not trivial to simulate by stacking together data with single type of anomalies as we do here.

The Conneau et al. task that most often shows the best transfer to our tasks (especially *Mod-Noun* and *Verb-Ob*) is *BShift*. This is sensible, given that that task involves detecting a switch in word order within a bigram. Given this similarity, we can expect to see some transfer from this task even in the absence of generalized anomaly encoding.
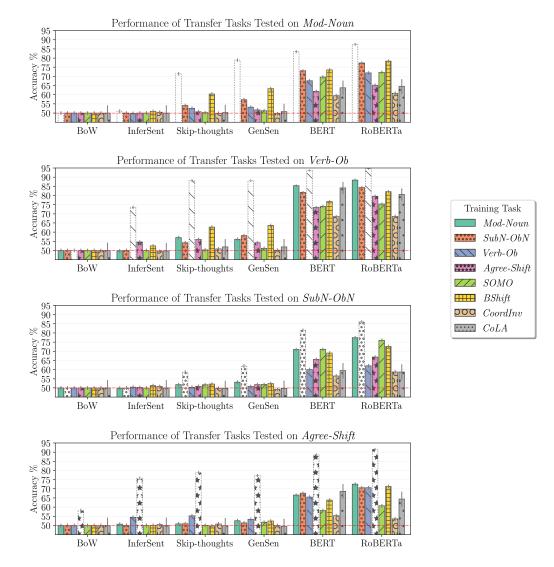
Figure 2: Results on transfer settings, by test tasks. Different patterns/colors represent different training tasks. Results for non-transfer settings (same training and test tasks) are shown in white bars.

As for how our generated anomaly types vary in supporting transfer to other anomaly types, we again note some differences between *Mod-Noun* and *Verb-Ob*. While *Verb-Ob* proved more accessible for detection than *Mod-Noun*, in the transfer setting we find that the broader hierarchical perturbation in *Verb-Ob* is often less conducive to transfer than *Mod-Noun*. Below we explore further to better understand what models are learning when trained on these anomalies.

7 Further analyses

7.1 Exploring false positives

The results above indicate that embeddings from these encoders contain non-trivial signal relevant for specific perturbations, but only BERT and RoBERTa show promise for encoding more general awareness of anomalies. To further explore the

anomaly signal learned from these representations, in this section we apply the learned anomaly classifiers to an entirely new dataset, for which no perturbations have been made. For this purpose, we use the *Subj-Num* dataset from Conneau et al. (2018).[7] By default we can assume that all sentences in these data are non-anomalous, so any sentences labeled as *perturbed* can be considered errors. After testing the pre-trained classifiers on embeddings of these unperturbed sentences, we examine these false positives to shed further light on what the classifiers have come to consider anomalies. We focus this analysis on BERT and RoBERTa, as the two models that show the best anomaly detection and the only signs of generalized anomaly encoding.

Error rates for this experiment are shown in Appendix Table 6. We see that in general the false pos-

---

[7]Dataset size is 10k sentences (test set).

Train task	Mod-Noun	Verb-Ob	SubN-ObN	Agree-Shift
BERT	4.9%	2.61%	4.81%	31.98%
RoBERTa	7.73%	3.6%	7.83%	22.13%

Table 1: Error rate on *Subj-Num* data.

itive rates for these models are very low. The highest error rates are found for the classifiers trained on *Agree-Shift*, and examination of these false positives suggests that the majority of these errors are driven by confusion in the face of past-tense verbs (past tense verbs do not inflect for number in English—so past tense verbs were few, and uninformative when present, in our *Agree-Shift* task). This type of error is less informative for our purposes, so we exclude the *Agree-Shift* classifier for these error analyses. For the other classifiers, the error rates are very low, suggesting that the signal picked up on by these classifiers is precise enough to minimize false positives in *normal* inputs.

To examine generality of the anomaly signal detected by the classifiers, we look first to sentences that receive false positives from multiple of the three reordering-based classifiers. We find that within the union of false positives identified across all three classifiers, sentences that are labeled as anomalous by at least two classifiers make up 28.6% and 35.6% for BERT and RoBERTa respectively—and sentences labeled as anomalous by all three classifiers make up 7.3% and 9.6%. Since no two classifiers were trained on the same perturbation, the existence of such overlap is consistent with some generality in the anomaly signal for the representations from these two models.

Table 2 lists samples of false positives identified by all three classifiers. While these sentences are generally grammatical, we see that many of them use somewhat convoluted structures—in many cases one can imagine a human requiring a second pass to parse these correctly. In some cases, as in the "Fireworks" example, there is not a full sentence—or in the "ornaments" example, there appears to be an actual ungrammaticality. The fact that the classifiers converge on sentences that do contain some structural oddity supports the notion that these classifiers may, on the basis of these models' embeddings, have picked up on somewhat of a legitimate concept of syntactic anomaly.

Of course, there are also many items that individual classifiers identify uniquely. We show examples of these in Appendix Table 5. The presence of such anomaly-specific errors is consistent with

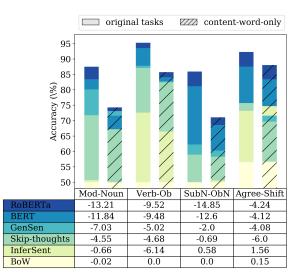|	Mod-Noun	Verb-Ob	SubN-ObN	Agree-Shift
RoBERTa	-13.21	-9.52	-14.85	-4.24
BERT	-11.84	-9.48	-12.6	-4.12
GenSen	-7.03	-5.02	-2.0	-4.08
Skip-thoughts	-4.55	-4.68	-0.69	-6.0
InferSent	-0.66	-6.14	0.58	1.56
BoW	-0.02	0.0	0.0	0.15

Figure 3: Performance of encoders on original tasks vs. "content-word-only" tasks (function words removed). Numbers in embedded table show change in accuracy from original to content-word-only setting.

our findings in Section 6, that even with BERT and RoBERTa the classifiers appear to benefit from some anomaly-specific signal in addition to the potential generalized anomaly signal.

Examining these classifier-specific false positives, we can see some patterns emerging. The *Mod-Noun* classifier seems to be fooled in some cases by instances in which a modifier comes at the end of a phrase (e.g., "a lovely misty gray"). For *Verb-Ob*, the classifier seems at times to be fooled by grammatical sentences ending with a verb, or by fronting of prepositional phrases. For *SubN-ObN*, the false positives often involve nouns that are likely uncommon as subjects, such as "bases". All of these patterns suggest that to some extent, the anomaly-specific cues that the classifiers detect are closely tied to the particulars of our perturbations—some of which may constitute artifacts—and in some cases, they raise the question of whether classifiers can succeed on these tasks based on fairly superficial word position cues, rather than syntax per se. We follow up on this question in the following section.

7.2 Role of content word order

To explore the possibility that classifiers may be succeeding in anomaly detection based on word position cues alone, rather than details of syntactic structure, we run a follow-up test using content-word-only versions of our probes. This serves as a test of how well the models can get by with coarser-grained information about content words positions.

BERT	Dolores asked pointing to a sway backed building made in part of logs and cover with a tin roof .
	Fireworks* , animals woven of fire and women dancing with flames .
	There were nice accessible veins there .
	Three rusty screws down and Beth spoke , making him jump .
	The pillars were still warming up , but the more powerful they got the more obvious it became to Mac about what was going on .
	The signals grew clearer , voices , at first faint , then very clear .
RoBERTa	One row , all the way across , formed words connected without spaces .
	And kidnappers with God only knew what agenda .
	The slums would burn , not stone nobleman keeps . "
	" Hull reinforcements are out of power .
	From inside that pyramid seventy centuries look out at us .
	The ornaments* she wore sparkled but isn 't noticeable much , as her blissful countenance shined over , surpassing it .

Table 2: Representative false positives shared by all three reordering classifiers.

Fig. 3 shows anomaly detection performance when embeddings reflect only content words, as compared to the original anomaly detection performance. We see that across tasks, anomaly detection performance of Skip-thoughts, GenSen, BERT, and RoBERTa are all reduced as a result of the loss of function words. BERT and RoBERTa in particular show substantial losses for the three reordering tasks, indicating that these models benefit significantly from function words for encoding the information that supports detection of these anomalies. It is also worth noting, however, that the models do retain a non-trivial portion of their original accuracy even with function words absent, supporting the idea that to some extent these perturbations can be detected through coarser position information rather than fine-grained syntax.[8] This is an observation worth keeping in mind, particularly when interpreting anomaly detection as evidence of syntactic encoding (e.g. Conneau et al., 2018).

8 Discussion

In the experiments above, we have taken a closer look at the nature of syntactic anomaly encoding in sentence embeddings. Using fine-grained variation in types of syntactic anomalies, we show differences in patterns of anomaly detection across encoders, suggesting corresponding differences in the types of anomaly information encoded by these models. While the margins of difference in anomaly-specific sensitivity are less dramatic between small RNN-based models and larger transformer models, when we examine the generality of the detected anomaly signal, we find that only BERT and RoBERTa show signs of higher-level

anomaly awareness, as evidenced by non-trivial transfer performance between anomalies.

What might be driving the anomaly encoding patterns indicated by our results? Explicit syntactic training does not appear to be necessary. GenSen is the only model that includes an explicit syntactic component in its training (constituency parsing), which could help to explain that model's comparatively strong performance on the individual anomaly detection tasks. However, it is noteworthy that GenSen performs mostly on par with Skip-thoughts, which constitutes just one of GenSen's objectives, and which uses only prediction of adjacent sentences. BERT and RoBERTa, the only models to show signs of more generalized anomaly encoding, have no explicit syntactic training at all. However, various findings have suggested that these types of models do develop syntactic sensitivity as a result of their more generalized training objectives (Goldberg, 2019; Liu et al., 2019a; Alleman et al., 2021; Tenney et al., 2019).

We can imagine various ways in which objectives involving prediction of words in context, as used by BERT and RoBERTa, could encourage learning of a generalized notion of syntactic anomaly. It may be the case that oddities occur together, in which case they could be mutually predictive and therefore of value for optimizing a prediction-based objective. More generally, anomalous sentences are likely identifiable as less probable, or more difficult to generate coherent predictions for. This relationship between anomaly and sentence probability raises a related question: Is this a problem for our conclusions here? Could models simply be identifying anomalous sentences as less probable, without any actual notion of syntactic anomaly? In NLP models, assessment of text probabilities is closely related to assessment of text naturalness and acceptability. For this reason, teasing apart general sensitivity to probability

---

[8]The notion that coarser position information alone can contribute non-trivially to anomaly identification is further supported by testing on *normal* sentences in *Subj-Num* when training on the content-word-only setting; for these results, see Appendix Table 6.

versus genuine awareness of syntactic grammaticality phenomena is a recurring challenge when testing syntactic knowledge in language models—and these things are similarly potentially entangled in our analyses here. To an extent this entanglement is inevitable and unproblematic: we necessarily expect syntactic anomalies to lower the probability of a string, and we can expect some awareness of syntactic anomaly to be important for assigning lower probability to an ungrammatical string. We can imagine, for instance, a situation in which a language model has no sensitivity to what constitutes good versus anomalous syntax, and thus assigns probability solely on the basis of word co-occurrence or other unrelated indicators of naturalness. In this sense, although it is not difficult to imagine how the close relationship between anomaly and sentence probability could be an explanation for findings that suggest anomaly awareness in these models, this does not change the fact that model representations may end up with genuine, detectable encoding of generalized anomaly information as a byproduct of probabilistic training—and this genuine anomaly encoding may be what we are detecting with our tests here. However, future work can examine further the relationship between syntactic anomaly and model perplexities, to explore whether these embeddings could show signs of anomaly sensitivity while in fact exclusively encoding confounding probabilistic information unrelated to syntactic anomaly.

Our content-word-only analysis provides one source of evidence on the relative importance of genuine syntactic sensitivity in success on our syntactic-related anomaly tasks. This test aims to tease apart the extent to which success on our tasks requires processing of finer-grained syntactic information, versus the extent to which models can succeed based on more superficial content word position information. We find in this analysis that most encoders do benefit from the finer-grained syntactic information provided by function words, supporting an important role for more advanced syntactic sensitivity in these tasks—however, we also find that substantial proportions of the observed detection accuracy can indeed be achieved with content words alone. To the best of our knowledge, we are the first to report this finding. This leaves us with two key takeaways. First, to an extent there is good reason to believe that a reasonable amount of genuine syntactic sensitivity is involved in the highest

levels of success on our anomaly tasks. Second, success on syntactic anomaly tasks can also be non-trivially inflated by use of more superficial cues. That is to say, as usual, these datasets have a habit of enabling learning from simpler cues than are intended. This takeaway highlights a need for caution in interpreting detection of reordering anomalies as evidence of deeper syntactic encoding per se (Conneau et al., 2018)—especially in probing datasets that use naturally-occurring data without exerting controls for confounding cues.

While these results have shed light on potential encoding of generalized syntactic anomaly knowledge in pre-trained models, there are many further questions to pursue with respect to these models' handling and understanding of such anomalies. We will leave for future work the problem of understanding in greater detail how model training may contribute to encoding of a generalized awareness of anomalies in text, how a genuine notion of syntactic anomaly could be further disentangled from general probability sensitivity, and how one could exploit models' awareness of anomaly for improving model robustness on downstream pipelines.

9   Conclusion

We have undertaken a direct study of anomaly encoding in sentence embeddings, finding impacts of hierarchical differences in anomaly type, but finding evidence of generalized anomaly encoding only in BERT and RoBERTa. Follow-up analyses support the conclusion that these embeddings encode a combination of generalized and anomaly-specific cues in these embeddings, with models appearing to leverage both finer-grained and coarser-grained information for anomaly detection. These results contribute to our understanding of the nature of encoding of linguistic input in embeddings from recent models. Future work can further explore the relationship between naturalness-oriented training and cultivation of abstract anomaly awareness, and how these insights can be leveraged for more robust and human-like processing of language inputs.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Matteo Alleman, Jonathan Mamou, Miguel A Del Rio, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2021. Syntactic perturbations reveal representational correlates of hierarchical phrase structure in pretrained language models. *arXiv preprint arXiv:2104.07578*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

Chomsky. 1957. *Syntactic Structures*. The Hague/Paris:Mouton.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics*, pages 133–144.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *ACL 2018*, volume 1, pages 2126–2136. Association for Computational Linguistics.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

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

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139.

Janet Dean Fodor, Weijia Ni, Stephen Crain, and Donald Shankweiler. 1996. Tasks and timing in the perception of linguistic anomaly. *Journal of Psycholinguistic Research*, 25(1):25–57.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42.

Tao Ge, Furu Wei, and Ming Zhou. 2018. Reaching human-level performance in automatic grammatical error correction: An empirical study. *arXiv preprint arXiv:1807.01270*.

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

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL-HLT*, pages 1195–1205.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. Bert & family eat word salad: Experiments with text understanding. *arXiv preprint arXiv:2101.03453*.

Homa B Hashemi and Rebecca Hwa. 2016. An evaluation of parser robustness for ungrammatical sentences. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1765–1774.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811.

Yair Lakretz, Cognitive Neuroimaging Unit, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in lstm language models. In *Proceedings of NAACL-HLT*, pages 11–20.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is bert surprised? layerwise detection of linguistic anomalies. *arXiv preprint arXiv:2105.07452*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

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

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

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448. Association for Computational Linguistics (ACL).

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

Thang M Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *arXiv preprint arXiv:2012.15180*.

Adam Poliak, Yonatan Belinkov, James Glass, and Benjamin Van Durme. 2018a. On the evaluation of semantic phenomena in neural machine translation using natural language inference. 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 513–523.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018b. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

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.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Fan Yin, Quanyu Long, Tao Meng, and Kai-Wei Chang. 2020. On the robustness of language encoders against grammatical errors. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

A Appendix

A.1 Cross-lingual encoders

Our probing tasks focus on anomalies defined relative to English syntax—but of course grammatical properties vary from language to language. Some of our perturbations produce constructions that are grammatical in other languages. We compare Universal Sentence Encoder (Cer et al., 2018) in the variants of both monolingual and cross-lingual (Chidambaram et al., 2019) trained models—Multi-task en-en, Multi-task en-fr, and Multi-task en-de. This allows us to examine impacts of cross-lingual learning, between English and different target languages, on anomaly sensitivity. [9]

From Table 3, we see that the two cross-lingual encoders (Multi-task en-fr and Multi-task en-de) do show slightly stronger anomaly detection relative to the monolingual model (Multi-task en-en) on *Mod-Noun*, *Verb-Ob*, and *Agree-Shift*, while having similar accuracy on *SubN-ObN*. This suggests that to accomplish the cross-lingual mapping

---

[9]We do not discuss this cross-lingual results in the main paper due to that the differences between mono-lingual encoder and its cross-lingual variants are small. But we do think the constant patterns behind the differences here are inspiring.

from English to French or English to German, these models may carry out somewhat more explicit encoding of syntactic ordering information, as well as morphological agreement information, resulting in encoded embeddings being more sensitive to corresponding anomalies relative to the monolingual model. As we have discussed in the main paper, anomaly detection in the *SubN-ObN* task likely involves understanding extra world knowledge, so it is perhaps not surprising that the cross-lingual component does not provide a boost in sensitivity on that task. For the most part, we find that the difference between the English-to-French and English-to-German mapping does not significantly impact encoding of the tested anomaly types.

(accuracy %)	Multi-task en-en	Multi-task en-fr	Multi-task en-de
Mod-Noun	54.858	57.539	59.109
Verb-Ob	63.204	66.188	66.345
SubN-ObN	56.91	56.372	56.641
Agree-Shift	56.125	61.746	61.33

Table 3: Anomaly detection results on mono-lingual encoder vs. its cross-lingual variants. (MLP)

As for the transferability of the anomaly encoding (especially on the multi-one transfer, see Table 7), we observe non-trivial performance improvement in the multi-task setting for the Multi-task en-en model, relative to the cross-lingual variants. This suggests that Multi-task en-en may result in a somewhat more generalized anomaly encoding, while cross-lingual variants are more sensitive to properties of individual anomalies.

A.2 Logistic regression results

			LR			
	BoW	InferSent	Skip-thoughts	GenSen	BERT	Roberta
Mod-Noun	50	57.595	69.329	79.459	82.903	87.189
Verb-Ob	50	72.078	85.001	87.469	93.033	95.187
SubN-ObN	50	56.451	57.774	61.723	80.48	85.484
Agree-Shift	55.34	72.403	72.93	72.538	83.778	91.676
			MLP			
	BoW	InferSent	Skip-thoughts	GenSen	BERT	Roberta
Mod-Noun	50	50.662	71.764	80.132	83.453	87.514
Verb-Ob	50	72.605	87.133	87.794	93.572	95.277
SubN-ObN	50	50	58.975	62.284	81.176	85.943
Agree-Shift	56.518	73.166	75.701	74.736	87.536	92.282

Table 4: Results (accuracy %) on original anomaly detection tasks, comparing between LR and MLP classifiers. The MLP results are the same as what has been shown in Fig. 1 in the main body.

As seen in Table 4, the results suggest that, for the most part, training with LR yields comparable

performance to that with MLP, consistent with the findings of Conneau et al. (2018).[10] We do find, however, that the LR classifier has higher accuracy for InferSent on *Mod-Noun* and *SubN-ObN*. This suggests that, to the extent that these anomalies are encoded (perhaps only weakly) in InferSent, they may in fact be better suited to linear extraction. For the task of *Agree-Shift*, most encoders show large improvements on MLP over LR, suggesting that morphological agreement anomalies are less conducive to linear extraction.

A.3 Error Analysis and the Role of Content Word Order

We show the error analysis results on the out-of-sample *Subj-Num* data in this part. Table 6 shows the error rate in terms of each training task (original and content-word-only), with the top two strongest encoders within our investigation. Table 5 lists sampled false positives identified exclusively by each classifier, along with corresponding initial observations.

As we see in the main paper, BERT and RoBERTa show non-trivial benefits from functional information for improving overall anomaly sensitivity, but the content-word-only setting can account for a substantial proportion of contribution for detecting the anomalies. This is roughly consistent with what we found in Table 6 that training from content word order only can lead to relatively lower error rate for most cases on normal sentences.

A.4 Transferring via Multi-Task Learning

Tables 7-8 list the multi-one transferring results with consistent training size compared to that of one-one transferring. We show the multi-one transfer results along with the performance change relative to the best one-one results obtained on the current test task when training with any one of the joint training tasks. Most of the multi-one transfer results show small amount of decrease from one-one transfer, suggesting that classifiers are still fitting to anomaly-specific properties that reduce transfer of anomaly detection.

A.5 Description of External Tasks for Transfer Training

SOMO *SOMO* distinguishes whether a randomly picked noun or verb was replaced with another

noun or verb in a sentence.

BShift *BShift* distinguishes whether two consecutive tokens within a sentence have been inverted.

CoordInv *CoordInv* distinguishes whether the order of two co-ordinated clausal conjoints within a sentence has been inverted.

CoLA *CoLA* tests detection of general linguistic acceptability in natural occurring corpus, using expert annotations by humans.

A.6 Description of Encoders

InferSent InferSent is a sentence encoder optimized for natural language inference (NLI), mainly focusing on capturing semantic reasoning information for general use.

Skip-thoughts Skip-thoughts is a sentence encoder framework trained on the Toronto BookCorpus, with an encoder-decoder architecture, to reconstruct sentences preceding and following an encoded sentence.

GenSen GenSen is a general-purpose sentence encoder trained via large-scale multi-task learning. Training objectives include Skip-thoughts, NLI, machine translation, and constituency parsing.

BERT BERT is a deep bidirectional transformer model, pre-trained on tasks of masked language modeling (MLM) and next-sentence prediction (NSP).

RoBERTa RoBERTa is a variant of BERT, and outperforms BERT on a suite of downstream tasks. RoBERTa builds on BERT's MLM strategy, removing BERT's NSP objective, with improved pretraining methodologies, such as dynamically masking. [11]

A.7 Implementation Details

Hyperparameters Dropout rate is set to be 0.25. Batch size is set to be 64. Early stopping is applied. The optimizer is Adam. The learning rate is explored within $\{0.01, 0.001, 0.0001, 0.00001\}$. The MLP classifier has one hidden layer of 512 units.

---

[10]Since we observe performance of LR to be mostly on par with one-hidden-layer MLP, we expect benefits of exploration with further classifier complexity to be limited.

[11]We use bert-large-uncased-whole-word-masking, and roberta-large. We take the average of fixed pre-trained embeddings from the last layer for all sentence tokens. Pilot experiments show comparable performance between the average of all token embeddings versus the first/CLS token embedding.

	Exclusive in *Mod-Noun*
BERT	The buildings here were all **a lovely misty gray** , which gave them a dreamlike quality . There are terrible slums in **London Daisy** , places you 'd never want to visit .
RoBERTa	Suddenly , his senses sharpened and he felt **less inebriated** . All the charts are in **drawers below the table** .
observation	Most of the samples involve a construction which is "seemingly a noun followed by a modifier format". For BERT, the samples seem to involve multiple adjectives in a row, where the final word is more frequently to be an adjective generally, and is more clear following the *Mod-Noun*-specific detection rules. For RoBERTa, e.g., "drawers below the table", "drawers" actually belongs to another prepositional phrase "in drawers" which is parallel to the followed by prepositional phrase "below the table".
	Exclusive in *Verb-Ob*
BERT	Slowly , the gatehouse **rose** . **Through their windows** , thick candles spread throughout flickered softly .
RoBERTa	**A satisfactory rate of exchange** I **feel** . The entire column stretched back almost as far as the eye could **see** .
observation	A clear pattern across both encoders: the error samples involve fronting such as prepositional phrase-fronting or object-fronting, or involve constructions end with a verb/verb phrase (thus not with a standard SVO structure).
	Exclusive in *SubN-ObN*
BERT	The **object** changed as he spoke . The **bases** were wide , and as the buildings climbed into the sky , they became narrower and branched off to connect to other buildings .
RoBERTa	That **joint** will help you sleep . The stealth **assassin** never belonged , but the reason will shatter his every conviction .
observation	For both encoders, the subject word of the sampled sentence is always an uncommon subject word.
	Exclusive in *Agree-Shift*
BERT	Not even the vendors who **stood** at their little shops or at their carts and called out their specials cared that I was there . The Tiger Man , still awake , **regarded** her with groggy eyes .
RoBERTa	My brows **went** up . A humorless laugh **escaped** his mouth and all I could do was stand mute , my heart breaking .
observation	Almost all of the error samples are with past tense main verb, across both encoders.

Table 5: Sampled examples for error analysis, along with some basic observed patterns. We list sampled typical examples for which the sentences are false positives exclusively in each of our four tasks. The bold text highlights words or constructions that possibly relate to what we think as cues that trigger our pre-trained classifier to predict the whole sentence as *perturbed*.

	original			
train task	*Mod-Noun*	*Verb-Ob*	*SubN-ObN*	*Agree-Shift*
BERT	4.9%	2.61%	4.81%	31.98%
RoBERTa	7.73%	3.6%	7.83%	22.13%
	content-word-only			
train task	*Mod-Noun*	*Verb-Ob*	*SubN-ObN*	*Agree-Shift*
BERT	1.82%	3.75%	2.0%	22.86%
RoBERTa	0.96%	1.64%	1.54%	4.92%

Table 6: Error rate on *Subj-Num* data, with a total size of 10,000 sentences (test set). The top rows show the results when the training on original tasks, while the bottom rows show the results when training on content-word-only tasks.

Probing data generation We use the premise sentences from the train, dev-matched, dev-mismatched datasets of MultiNLI, with repeats discarded according to the promptID. [12] [13]

We adopt an approach that we refer to as "exhaustive" perturbation: modifying all instances of a given structure within a sentence, to ensure that sentences have internal structural consistency—e.g., a perturbed sentence in *Mod-Noun* will not contain both "modifier+noun" and "noun+modifier" structures—thus avoiding inconsistency serving as an extra signal for detection. [14]

For each task, we use training data of 71k sentences, and dev and test data of 8.9k sentences.

For the *SubN-ObN* and the *Mod-Noun* tasks, [15]

---

[12] The following tools are used for our generation: nltk.tree https://www.nltk.org/_modules/nltk/tree.html, and spaCy https://spacy.io. The transformation tool (along with extra rules) for *Agree-Shift*: https://www.clips.uantwerpen.be/pages/pattern.

[13] The ratio of plural verbs to single verbs (VBP/VBZ) in the original sentences is 1.044/1.

[14] The average number of modifications per sentence is 1.69, 1.97, 1.0, and 1.97 for *Mod-Noun*, *Verb-Ob*, *SubN-ObN*, and *Agree-Shift*), respectively. Note that when we instead restrict perturbations to a single modification per sentence, we see that the same basic patterns across tasks are retained.

[15] For the task of *Mod-Noun*, this could in particular happen

Encoder/Train task	multi-task (Verb-Ob + SubN-ObN + Agree-Shift)		multi-task (Mod-Noun + SubN-ObN + Agree-Shift)		multi-task (Mod-Noun + Verb-Ob + Agree-Shift)		multi-task (Mod-Noun + Verb-Ob + SubN-ObN)	
	Mod-Noun	Δ	*Verb-Ob*	Δ	*SubN-ObN*	Δ	*Agree-Shift*	Δ
BoW	50.0	0.0	50.0	0.0	50.0	0.0	50.0	0.0
InferSent	50.011	**0.011**	54.61	**0.012**	51.369	**0.92**	54.337	-0.1
Skip-thoughts	53.298	-0.83	63.022	**6.001**	51.357	-0.46	55.289	**0.034**
GenSen	55.531	-1.772	60.61	**2.456**	52.984	-0.213	54.729	**1.401**
BERT	73.996	-0.101	83.614	-0.919	72.616	**0.09**	71.257	**0.941**
Multi-task en-en	53.455	**0.123**	54.587	-2.804	52.984	-0.247	51.154	**0.258**
Multi-task en-fr	52.165	-0.797	53.41	-0.28	52.244	-1.054	52.824	-0.258
Multi-task en-de	52.008	-0.381	54.419	**0.179**	52.333	-0.73	52.387	-0.863

Table 7: Transfer results of multi-one transferring among our generated tasks with consistent training size. The amount of training size of multi-task training is consistent with one-one transferring. The columns of Δ show how much the multi-one transferring improves or drops from the best one-one result. The improvements (positive Δ values) are bolded.

Encoder/Train task	multi-task (*SOMO + BShift + CoordInv*)							
	Mod-Noun		*Verb-Ob*		*SubN-ObN*		*Agree-Shift*	
	acc	Δ	acc	Δ	acc	Δ	acc	Δ
BOW	50.0	0.0	50.033	**0.033**	50.017	**0.017**	50.0	0.0
Infersen	50.067	-0.833	50.217	-2.316	50.1	-1.2	49.933	-0.5
Skip-thoughts	57.767	-2.666	57.733	-5.05	51.3	-0.85	50.383	-0.4
GenSen	58.533	-4.75	58.867	-4.75	50.533	-1.834	50.783	-1.617
BERT	69.883	-3.967	73.683	-5.1	69.317	-0.75	62.367	-1.55
Multi-task en-en	53.283	**0.283**	54.0	-0.533	52.617	**0.134**	51.017	**0.084**
Multi-task en-fr	50.767	-0.933	52.5	-5.467	51.85	-0.45	51.05	**0.117**
Multi-task en-de	49.7	-3.4	50.167	-5.3	50.3	-1.167	50.033	-0.5

Table 8: Transfer tasks jointly trained on multi-task learning with all of Conneau et al. tasks, tested on each of our generated tasks, with consistent training size.

sometimes the case might arise that the resulting perturbed sentences are still normal or acceptable, but perhaps somewhat stranger or less probable to occur in the wild, e.g., "man bites dog". However, this should be a rare case, as the original sentences are long enough to involve adequate context to distinguish normal from perturbed examples.

---

with noun phrases involving noun-noun compounds.