Implicit representations of event properties within contextual language models: Searching for "causativity neurons"

Esther Seyffarth[†] **Younes Samih**[†] **Laura Kallmeyer**[†] **Hassan Sajjad**[‡]
[†]Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[‡]Qatar Computing Research Institute, Hamad Bin Khalifa University
{seyffarth,samih,kallmeyer}@phil.hhu.de hsajjad@hbku.edu.qa

Abstract

This paper addresses the question to which extent neural contextual language models such as BERT implicitly represent complex semantic properties. More concretely, the paper shows that the neuron activations obtained from processing an English sentence provide discriminative features for predicting the (non-)causativity of the event denoted by the verb in a simple linear classifier. A layer-wise analysis reveals that the relevant properties are mostly learned in the higher layers. Moreover, further experiments show that appr. 10% of the neuron activations are enough to already predict causativity with a relatively high accuracy.[1]

1 Introduction and motivation

In natural language processing (NLP), machine learning models based on artificial neural networks have achieved impressive results in recent years, due to large amounts of available training data and powerful computing infrastructures. Contextual language models (LMs) such as ELMO (Peters et al., 2018), BERT (Devlin et al., 2019), and XLNet (Yang et al., 2019) have particularly contributed to this. However, it is oftentimes not clear which kinds of generalizations these models make, i.e., what exactly they learn. In this respect, neural networks suffer from a lack of transparency and interpretability. Recent research has started to investigate these questions. Since the successful use of neural word embeddings and LMs (e.g., Word2Vec, Mikolov et al. 2013; ELMO, Peters et al. 2018; BERT, Devlin et al. 2019) for a range of NLP/NLU tasks, it is clear that LMs capture meaning to a certain degree, in particular lexical meaning. Concerning syntactic information, work on different types of language models, in particular RNNs and transformer-based contextual language models, has shown that these models learn morphology (Liu et al., 2019a), syntactic structure and syntactic preferences to a certain degree (see Futrell and Levy, 2019; Lin et al., 2019; Hewitt and Manning, 2019; McCoy et al., 2020; Wilcox et al., 2019; Hu et al., 2020; Warstadt et al., 2020).

In this paper, we expand the question of what linguistic properties these models learn towards whether pretrained contextualized models capture more abstract semantic properties, in particular properties that contribute to the structure of the semantic representation underlying a given sentence. More concretely, we investigate whether an LM such as BERT represents whether a sentence denotes a causative event or not. If this was the case, we would expect a systematic difference between for instance BERT's neuron activations for (1-a) and for (1-b).

(1) a. Kim broke the window.
b. Kim ate an apple.

Note that the two sentences share almost no lexical elements, so the neuron activations are expected to be mostly different. Our research question is focused on whether there are systematic activation patterns that can be observed that are common to all instances of causative sentences, and others that are common to all instances of noncausative sentences, independent of sentence content.

One of the common approaches to probe neural network models is to use a probing classifier. Given a linguistic property of interest, the idea is to extract contextualized activations of units (words/phrases/sentences) relevant to the property. A classifier is then trained to learn the property by using the extracted activations as features. The performance of the classifier is taken to approximate

[1]Our datasets are available at https://github.com/eseyffarth/predicting-causativity-iwcs-2021

Proceedings of the 14th International Conference on Computational Semantics, pages 110–120
June 17–18, 2021.

the degree to which the language model learned the linguistic property. We also use probing classifiers and probe the model as a whole, its individual layers and its neurons with respect to causativity. We use the NeuroX toolkit (Dalvi et al., 2019b) to conduct the probing experiments.

We experiment using two 12-layer pretrained models, BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019), as well as a distilled version of BERT, DistilBERT (Sanh et al., 2019). Our findings and contributions are as follows: We create a novel dataset of sentences with verbs that are labeled for causativity/non-causativity. Using this dataset for probing, we show that this abstract semantic property is learned by the pretrained models. It is better represented in the higher layers of the model and, furthermore, there is a subset of appr. 10% of the neurons that encodes the property in question.

2 Related work

A number of interpretation studies have analyzed representations of pre-trained models and showed that they learn linguistic information such as part of speech tagging, semantic tagging and CCG tagging (Conneau et al., 2018; Liu et al., 2019a; Tenney et al., 2019a,b; Voita et al., 2019). A typical procedure to analyze representation is a post-hoc analysis using a probing classifier. It has been shown that word-level concepts are learned at lower layers while sentence-level concepts are learned at higher layers (Liu et al., 2019b). Dalvi et al. (2019a) extended the layer-level analysis towards individual neurons of the network. They proposed linguistic correlation analysis (`LCA`) to identify neurons with respect to a linguistic property. Durrani et al. (2020); Dalvi et al. (2020) later used `LCA` to analyze pre-trained models in the context of linguistic learning and redundancy in the network respectively.

In this work, we also aim to analyze pre-trained models at model-, layer- and neuron-level using post-hoc analysis methods. Different from others, we concentrate on an abstract, structure-building semantic property, namely causativity of events. Our focus is on *lexical causatives*, that is, verbs whose lexical meaning has a causative aspect (Dowty, 1979). In Dowty's aspect calculus, such verbs are analyzed as $[\phi$ CAUSE $\psi]$, where ϕ and ψ are sentences and causation is a "two-place sentential connective", notably even for sentences that only contain a single verb phrase. Thus, *John killed Bill* is decomposed as in (2) (Dowty, 1979, p. 91).

(2) [[*John does something*] CAUSE [BECOME¬[*Bill is alive*]]]

The "semantically bipartite" nature of causative verbs means that sentences with such verbs actually express not one event, but two subevents, one being the causing event and the other one being the caused event, or result, of the first. This event structure is a challenge to model with NLP systems when no superficial indicators for causativity are available. While there are verbs that are lexically causative (such as *refresh*) and verbs that are lexically noncausative (such as *prefer*), there are also verbs that vary in their causativity depending on the context in which they appear (such as *open*). Our goal is to determine to what extent the causativity or noncausativity of these types of verbs is implicitly learned by large language models.

3 Method

Over the last years, there has been an increasing interest in assessing linguistic properties encoded in neural representations. A common method to reveal these linguistic representations employs diagnostic classifiers or probes (Hupkes et al., 2018). A common diagnostic classifier is a linear classifier trained for the underlying linguistic task, using the activations generated from the trained neural network model as features. The performance of the classifier is used as a proxy to measure the amount of linguistic information present in the activations. We also use a linear classifier for probing.

Consider a pre-trained neural network model $\mathbf{M}$ with L layers: $\{l_1, l_2, \ldots, l_L\}$, where each layer l_i is of size H. Given a dataset $\mathbb{D} = \{s_1, s_2, ..., s_T\}$ consisting of T sentences, the contextualized embedding of sentence s_j at layer l_i is $\boldsymbol{z}_j^i = l_i(s_j)$. In pretrained models like BERT, a special token [CLS] is appended with every training instance during training. The token is later optimized for sentence embedding during transfer learning (Devlin et al., 2019). We consider the representations of [CLS] for sentence embedding in this study. The [CLS] representation extracted from various layers is used as input features to the probing classifier.

Model-level probing: To assess to what extent a linguistic property is learned in the model, we first take the sentence representations of all layers as features for linear classification, i.e., all $\boldsymbol{z}_j^i$ for

$1 \leq i \leq L$ and $1 \leq j \leq H$. The classifier is trained by minimizing the following loss function:

$$\mathcal{L}(\theta) = -\sum_{j} \log P_\theta(t_{s_j}|s_j) \quad (1)$$

where t_{s_j} is the predicted label for sentence s_j. In this work, binary labels are used to encode whether the property is present in a sentence or not.

Layer-level probing: Here, we question how much individual layers of a model represent our property of interest. We train a linear classifier on the activations of each individual layer. The performance of each layer serves as a proxy to how much information it encodes with respect to our property.

Neuron-level probing: While the layer-level probing tells about how much linguistic information is learned in a layer, it does not tell about the learning of individual neurons in the network. It is possible that while a particular layer performs best in the layer-level probing, the best neurons learning about the linguistic property are spread across many layers. In neuron-level probing, we aim to identify the most salient neurons across the network that learn the linguistic property at hand.

We follow the linguistic correlation analysis method (`LCA`) of Dalvi et al. (2019a) to conduct this analysis. Given representations of the model as in the model-level probing, `LCA` trains an ElasticNet (Zou and Hastie, 2005) classifier, and provides a salient list of neurons with respect to the linguistic property. ElasticNet provides a balance between selecting very focused localized features and distributed features (here: neurons). Equation (2) gives the loss function:

$$\begin{aligned}\mathcal{L}(\theta) \quad = \quad & -\textstyle\sum_j \log P_\theta(t_{s_j}|s_j) \\ & +\lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2^2\end{aligned} \quad (2)$$

where λ_1 and λ_2 are parameters, for which we use the suggested value of 0.00001 (Dalvi et al., 2019a).

4 Data

To prepare our datasets, we create different sets of verbs that are labeled for (non)causativity, and then use them as seeds to collect sentences from a corpus to be used as input to the classifier.

4.1 Verb set selection

Causative and noncausative verbs We collect a set of English verbs that are either always causative or never causative when appearing in basic transitive sentences (NP V NP). This property is derived from VerbNet 3.3 (Kipper et al., 2000) according to the event-semantic description of each basic transitive syntactic frame in each verb class. We only consider members of VerbNet classes where either all basic transitive frames or none of them are associated with causativity. Two trained linguists manually prune the lists of causative and noncausative verbs to remove ambiguous verbs and other edge cases. This results in a list of 2157 causative and 617 noncausative verbs.

Alternating verbs We also create a set of verbs whose causativity property depends on whether they appear in transitive or intransitive sentences. This is the case for verbs in VerbNet that are marked with the "Causative" property in basic transitive syntactic frames, and with the "Inchoative" property in basic intransitive frames. These verbs participate in the causative-inchoative alternation. They represent a special case for our experiments because the classifier needs to distinguish between causative and noncausative uses of identical verbs, whereas the sets of causative and noncausative verbs are completely distinct. In this setting, the classifier cannot rely purely on the verb lemma (because alternating verbs can appear in both classes), and it also cannot rely purely on the (in)transitivity of sentences (because verbs outside the alternation can be causative in intransitive sentences). Since this makes the task more difficult, we expect the classification accuracy to be lower in this setting than in settings with non-alternating verbs.

4.2 Sentence selection

We collect three datasets for our experiments.[2] All sentences are extracted from ENCOW (Schäfer and Bildhauer, 2012; Schäfer, 2015), an English web corpus (9.6 billion tokens) annotated with dependencies created with MaltParser. Each dataset contains 40,000 sentences in the `train` portion, 5,000 sentences in `dev` and 5,000 sentences in the `test` portion. Each portion contains an equal number of causative and noncausative instances. Each test set contains sentences that were not previously seen in the train set, but not all verbs in the test set are unseen.

[2] All datasets are available at https://github.com/eseyffarth/predicting-causativity-iwcs-2021

Transitive sentences, same sentence length The first dataset (D_{tr_5}) is based on the sets of causative and noncausative verbs and contains only transitive sentences of length 5 (including punctuation). This yields a dataset where all sentences have the same basic syntactic pattern. Examples are given in (3) (root verbs in bold).

(3) a. The answer **surprised** me . *(caus)*
b. It **contains** no surprises . *(noncaus)*

Transitive sentences, varying sentence length The second dataset (D_{tr}) is based on the same verb sets, but contains sentences of varying lengths between 5 and 20 tokens. Examples are given in (4).

(4) a. This **affects** the calculation . *(caus)*
b. I **envy** you in that respect ! *(noncaus)*

Intransitive and transitive sentences, varying length The third set (D_{all}) is based on the verb set that includes verbs in the causative-inchoative alternation. Sentences in D_{all} are either transitive or intransitive and have a length between 5 and 20 tokens. Again, each portion contains an equal number of causative and noncausative instances, consisting of verbs of all three types (alternating, always causative, always noncausative). Examples are given in (5); note that (5-e) and (5-f) share the same alternating root verb.

(5) a. I **bring** a book ! *(caus)*
b. Everything about them **intimidates** . *(caus)*
c. Each layer **had** its own opacity . *(noncaus)*
d. A total of 24 people **attended** . *(noncaus)*
e. He **opened** the pack . *(caus)*
f. The main console **opens** . *(noncaus)*

5 Evaluation

5.1 Experimental Settings

Pre-trained models We conduct experiments using three transformer-based pre-trained language models: BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and XLNet (Yang et al., 2019). The BERT model is an auto-encoder trained with two unsupervised objectives: masked word prediction and next sentence prediction. It is pre-trained on Wikipedia text and BooksCorpus (Zhu et al., 2015), and comes with hundreds of millions of parameters. It contains an encoder with 12 Transformer blocks, hidden size of 768, and 12 self-attention heads. DistilBERT is an approximate distilled version of BERT. It is comprised of 6 encoder layers while retaining 97% of BERT performance. We also employ XLNet-base in all our experiments. Although it is trained with the same parameter configurations as BERT-base, it uses improved training methodology based on a permutation auto-regressive objective function.

Data	BERT	DistilBERT	XLNet
D_{tr_5}	95.24	93.34	90.92
D_{tr}	89.48	87.28	88.84
D_{all}	85.28	83.96	86.00

Table 1: Model-level results (accuracy) using all neurons for classification

Since we are interested in analyzing sentence representations, we use the representation of the [CLS] token. However, the representation of [CLS] is not optimized for sentence embedding in the pre-trained models. In order to tune it for sentence representation, we fine-tune the pre-trained model on a sentence classification task, the Stanford sentiment treebank (Socher et al., 2013). We understand that by fine-tuning the pre-trained model, the representations of the network are tuned for the task. An alternate strategy is to use average activations of words in a sentence as sentence representation. We did not explore it in this paper.

Probing Classifier We train a linear classifier using a categorical cross-entropy loss, optimized using Adam. For neuron-level analysis, we used elastic-net regularization. We used the recommended values of elastic-net parameters, i.e., λ_1 and λ_2 each equal to 0.0001.

5.2 Results

Model-level Results Table 1 presents the results of using all neuron activations of the model as features for classification. The general high classification results show that the model has learned causitivity. However, as the dataset becomes hard in terms of varying sentence length and including more challenging instances with alternating verbs, the performance drops to as low as 83.96% for DistilBERT, which is still substantially better than random performance (50%).

Layer-level Results Here we want to see which layers of pretrained models learn causativity. We train our probing classifier on individual layers. Figure 1 summarizes the results. As a general trend, causitivity is best represented at the higher layers

		BERT	DistilBERT	XLNet
	Neu_a	9984	5372	9984
D_{tr_5}	Neu_t	1000/10%	540/10%	300/3%
	Acc_t	95.06	92.6	92.02
D_{tr}	Neu_t	1000/10%	540/10%	1000/10%
	Acc_t	88.70	86.06	89.24
D_{all}	Neu_t	1000/10%	540/10%	1000/10%
	Acc_t	86.48	82.66	86.8

Table 2: Selecting minimal number of neurons. Neu_a = Total number of neurons, Neu_t = Top selected neurons, Acc_t = Accuracy after retraining the classifier using only selected neurons.

	verb type	BERT	DistilBERT	XLNet
D_{tr_5}	caus	95.24	93.04	98.84
	noncaus	96.44	94.92	84.12
D_{tr}	caus	90.44	89.60	90.44
	noncaus	88.88	84.76	86.12
D_{all}	all alternating	81.52	75.43	83.05
	alt. caus	89.73	84.35	94.87
	alt. noncaus	52.59	43.97	41.38
	nonalt. caus	91.25	84.60	93.93
	nonalt. noncaus	85.36	86.03	79.28

Table 3: Accuracy per verb type and data set in all settings. D_{tr_5}, D_{tr} and D_{all} each contain an equal number of caus(ative) and noncaus(ative) instances.

of the models, which is in line with previous findings that sentence-level properties such as syntax are better learned at higher layers (Durrani et al., 2020). For all models, we see a slight drop in the performance for the last layer, which is due to the fact that the last layer is optimized for the objective function (Kovaleva et al., 2019). Compared to BERT and DistilBERT, the middle layer of XLNet consistently showed a small drop in the performance for all datasets. This trend is more prevalent in the neuron-level results. We discuss it later in this section.

Neuron-level Results We use LCA to determine a minimal set of neurons that still achieve a classification performance (Acc_t) within 2% of the performance using all the neurons of the network for classification. We additionally evaluate the effectiveness of the LCA method by comparing the classification performance using the top selected neurons with the randomly selected neurons. We found the salient neurons of LCA to perform substantially better than random neurons.

Table 2 presents the numbers of salient neurons selected for each model and for each dataset together with the resulting classification accuracy. Note that in the case of BERT and the dataset D_{all} and also for XLNet on all datasets, the accuracy increased due to the elimination of non-discriminative features.

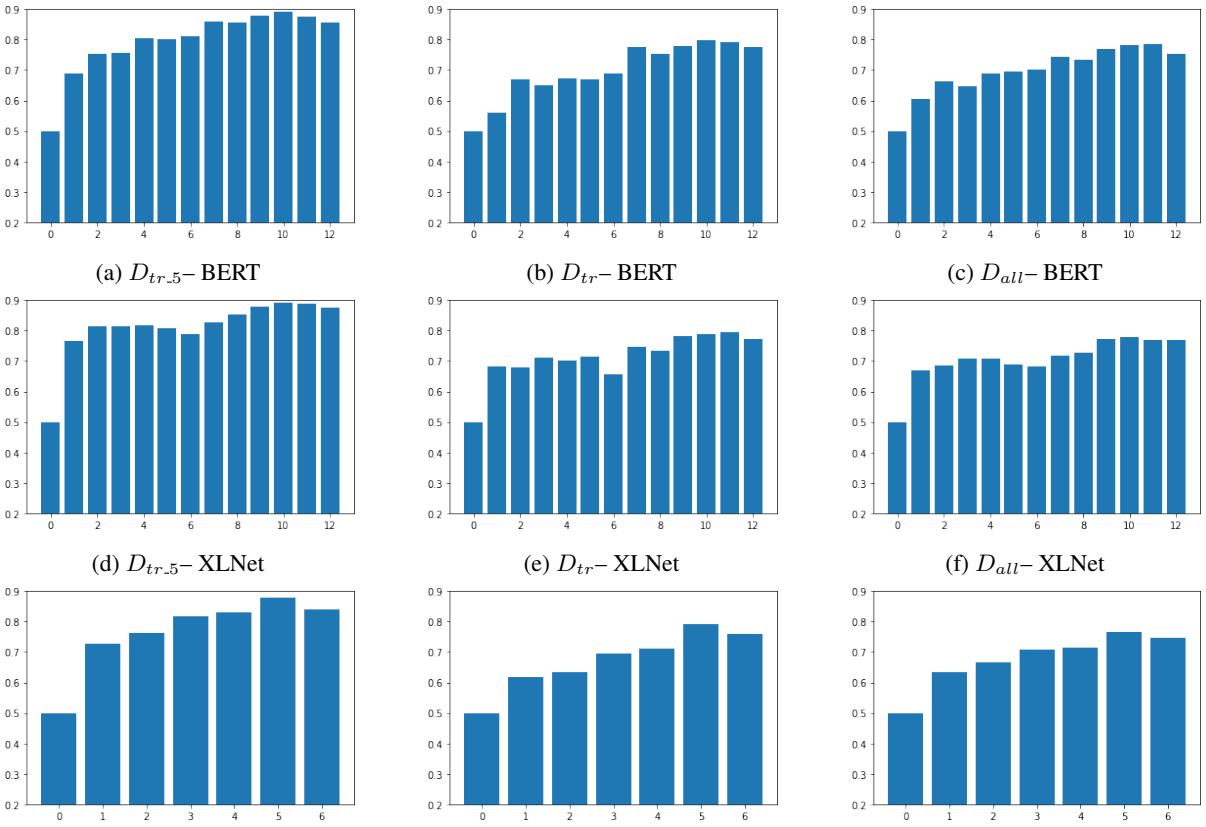
Given salient neurons with respect to our task, we observe their distribution across the model. Figure 2 summarizes the results. Across all models and datasets, the LCA method never selected any neurons from the embedding layer. This is in line with the layer-wise results where the performance using embedding layer representation is similar to random classification, i.e., no causativity information is present.

For BERT and DistilBERT, the distribution of salient neurons is skewed towards higher layers (excluding top layer), i.e., causativity information is more represented at the higher layers. XLNet presents a slightly different picture where the salient neurons selected from the middle layers are substantially lower than most of the other layers. As the task becomes harder, the contribution of lower middle layers (3-4) substantially increases while the last layer contribution drops.

The number of neurons selected from middle layers (5-6 in the case of 12 layer models and 3 in the case of 6 layer models) are substantially lower than the neighbouring layers across all models and data sets. We hypothesize that learning causitivity requires word-level and sentence-level information which is dominating at the lower and higher layers.

6 Discussion

As shown in Table 1, all classifiers performed best on D_{tr_5}. With little syntactic variation between instances in D_{tr_5}, this is the least challenging setting for the task: The verbs and arguments in each sentence are the main indicators for the classifiers to identify causativity. In D_{tr}, all models achieve slightly lower accuracy. Longer sentences are more likely to contain conjunctions or subordinate clauses, which may distract the classifiers from the sentence's (non)causative root verb and its arguments. As expected, the lowest accuracy scores are observed in D_{all}, which includes both transitive and intransitive sentences, as well as alternating verbs whose causativity property changes in these different environments. Table 3 shows that all three models mislabel alternating verbs more often than nonalternating verbs. BERT and XLNet



(a) D_{tr_5}– BERT (b) D_{tr}– BERT (c) D_{all}– BERT

(d) D_{tr_5}– XLNet (e) D_{tr}– XLNet (f) D_{all}– XLNet

(g) D_{tr_5}– DistilBERT (h) D_{tr}– DistilBERT (i) D_{all}– DistilBERT

Figure 1: Layer-wise results: X-axis = Layer number, Y-axis = Classification accuracy

achieved the best accuracy for causative verbs in almost all experiments, while DistilBERT often performed better on noncausative verbs.

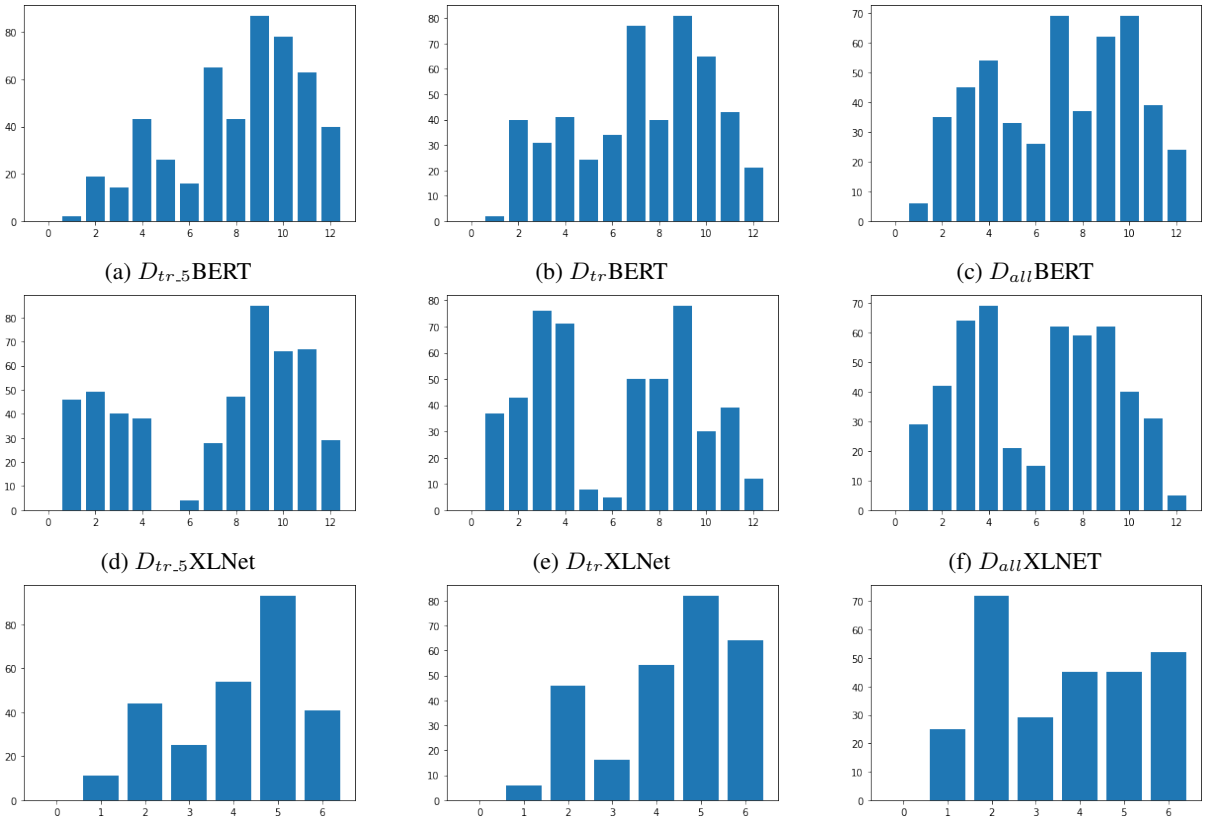
Our datasets are randomly collected from a larger corpus with no regard for verb frequency. This results in datasets where some verbs occur only once or twice, some are never seen in the training data, and some are more common. Our goal is to determine whether the classifiers successfully learn to predict (non)causativity, independently of specific verb lemmas. The results reported so far are all averaged over all verbs in a dataset, illustrating that some models are more successful on the classification task than others (e.g. BERT achieving higher accuracy scores than the other models on the first two datasets). Additionally, it is also worth exploring the accuracy of the classifiers for individual verbs, particularly those that are most likely to be mislabeled by any of the classifiers. Table 4 reports the two most-mislabeled verbs of each type per dataset (across all models). Notably, the XLnet classifier consistently makes more mistakes with noncausative instances than with causative ones, as is also apparent from Table 3.

Broadly, the frequently mislabeled verbs fall in three categories: 1. presumed errors due to parsing mistakes and subsequent errors in the gold data; 2. errors due to incorrect labels of ambiguous verbs in the gold data; 3. errors due to an ambiguity between full verb, light verb, and auxiliary verb.

Presumed errors due to parsing mistakes and subsequent errors in the gold data Most of the frequently-mislabeled verbs in D_{tr_5} fall into this category. These verbs occur only a few times each, indicating that they do not represent a deeper structural issue with the classifiers; for instance, sentences with the root verb *mark* occasionally appear incomplete in ENCOW, as exemplified in (6).

(6) the symptoms marked gr . (ENCOW-02-23709973)

The verb *sound* is labeled as a causative verb in our gold data (e.g., "to sound the bells"), but appears often in another word sense, as exemplified in (7-a). In these sentences, the verb does not have a direct object as expected; the reason for their inclusion in our datasets is an incorrect dependency parse in ENCOW. In other words, the causative



(a) D_{tr_5}BERT (b) D_{tr}BERT (c) D_{all}BERT

(d) D_{tr_5}XLNet (e) D_{tr}XLNet (f) D_{all}XLNET

(g) D_{tr_5}DistilBERT (h) D_{tr}DistilBERT (i) D_{all}DistilBERT

Figure 2: How top neurons spread across different layers for each causativity dataset. X-axis = Layer number, Y-axis = Number of neurons selected from that layer

gold label is assigned by mistake. A similar case is *mean*; as with *sound*, many instances do not involve a direct object at all, as exemplified in (7-b), but are included because of an incorrect parse.

(7) a. that sounds so scary !!! (ENCOW-05-11095175)
b. you mean screw justice ? (ENCOW-14-01839826)

D_{all} also contains incorrect gold labels that are to a large extent due to parsing errors, for instance *bring*. All sentences included in (8) were parsed as having *bring* as their root verb. That the classifiers tended to assign a noncausative label to these sentences suggests that they instead assigned labels for *take for granted*, *love*, or *be*, respectively (which is actually correct).

(8) a. people take for granted what tax money brings . (ENCOW-11-16881058)
b. knowledge is power , and what americans really love is the power knowledge brings . (ENCOW-13-11898010)
c. sugar is a barrow boy with all that epithet brings . (ENCOW-10-21805613)

In future work, we will improve our datasets to minimize the number of this type of errors, using a more recent dependency parser and some manual checking.

Errors due to incorrect gold labels of ambiguous verbs In D_{tr}, *face* is the most mislabeled causative verb. The presumed causative label for this verb comes from the VN class *confront-98*, which contains verbs such as *target* or *combat*. However, the mislabeled examples from the dataset seem to evoke a weaker, more passive sense of *face*, as in (9-a), where human annotators might not assign a causative label. In these cases, the label assigned by the classifier is actually correct, while the gold label is not. The mislabeled instances of *cover* in D_{all} are, similarly to *face*, an artefact of verb polysemy and should in fact not be regarded as causative sentences, as exemplified in (9-b).

(9) a. older mums face similar risks . (ENCOW-05-25724129)

	BERT	DistilBERT	XLNet
causative verbs in D_{tr_5}			
mark	6 (60.00%)	6 (60.00%)	2 (20.00%)
sound	1 (2.00%)	9 (18.00%)	4 (8.00%)
noncausative verbs in D_{tr_5}			
leave	4 (10.00%)	7 (17.50%)	15 (37.50%)
mean	1 (1.37%)	2 (2.74%)	15 (20.55%)
causative verbs in D_{tr}			
face	11 (25.00%)	11 (25.00%)	17 (38.64%)
express	12 (33.33%)	8 (22.22%)	10 (27.78%)
noncausative verbs in D_{tr}			
leave	7 (17.07%)	19 (46.34%)	17 (41.46%)
represent	8 (8.42%)	17 (17.89%)	15 (15.79%)
alternating causative verbs in D_{all}			
set	3 (8.82%)	10 (29.41%)	1 (2.94%)
open	3 (12.00%)	4 (16.00%)	3 (12.00%)
alternating noncausative verbs in D_{all}			
close	4 (57.14%)	5 (71.43%)	6 (85.71%)
open	4 (66.67%)	2 (33.33%)	5 (83.33%)
nonalternating causative verbs in D_{all}			
cover	9 (6.52%)	32 (23.19%)	10 (7.25%)
bring	9 (9.47%)	10 (10.53%)	9 (9.47%)
nonalternating noncausative verbs in D_{all}			
have	25 (4.64%)	19 (3.53%)	43 (7.98%)
be	20 (10.81%)	25 (13.51%)	36 (19.46%)

Table 4: Most mislabeled verbs in all settings. Each cell states the number of instances with the given verb with an incorrect label, giving the absolute number followed by the percentage of all instances with this verb.

b. the manual that comes with the game covers everything you need to know , including the mission editor . (ENCOW-08-06019647)

Sentences with the verb *represent* are frequently labeled as causative by one or more of the classifiers. When the verb is used in a legal or political sense, as in (10), this may in fact be appropriate. Since our verb sets are labeled on the lemma level and we do not perform any word sense disambiguation, these differences are not explicitly marked in our datasets, so these sentences are counted as mislabeled instances.

(10) they represent the voice of over 80,000 students and 62,000 members in 155 countries . (ENCOW-09-01862399)

In D_{tr}, all classifiers occasionally label instances of noncausative *leave* as causative, particularly XLNet. *leave* is a member of the VN classes *become-109.1-1-1*, *escape-51.1-1-1*, *fulfilling-13.4.1*, *future_having-13.3*, *keep-15.2*, and others. While not all of these classes license basic intransitive sentences of the type included in our datasets, this illustrates the polysemy of *leave*, which might be an explanation for the relatively high number of mislabeled instances in our experiments.

Generally, in D_{all}, noncausative alternating verbs are among the most mislabeled verbs. Since the dataset contains different numbers of verbs of each type, this may be a sparsity effect more than an effect of these verbs being more difficult to label. This question will be approached with new datasets in future work.

The reason for most errors of this type is that our datasets were created automatically with the help of a lexical resource. In order to avoid such polysemy issues, a version of the datasets with human annotations would be necessary.

Errors due to an ambiguity between full verb, light verb, and auxiliary verb Finally, the verbs *have* and *be* are the most mislabeled nonalternating noncausative verbs in D_{all}. These verbs appear in light verb constructions, as auxiliary verbs, and in a range of word senses that can be causative or noncausative. The examples in (11) illustrate why the classifiers are struggling to label such sentences as noncausative. Note that in all cases, the MaltParser annotations provided alongside ENCOW mark a form of *have* as the root verb.

(11) a. hi we have just moved house and the house has no tv aerial . (ENCOW-11-17855426)
b. we had a small cup made up not long ago with a very simple design . (ENCOW-06-00570494)
c. local people have the power to stop this by not buying counterfeit products . (ENCOW-08-19775040)

ENCOW was parsed between 2015 and 2018 using the standard *engmalt* model available on the MaltParser website (Roland Schäfer, p.c.) This type of error would be minimized if a more recent dependency parser was used.

To summarize, many of the "errors" of the classifiers are actually not errors but incorrect labels in the gold data. This means that the classifiers might be better in predicting causativity than assessed by our evaluation.

7 Conclusion

We set up a series of classification experiments with a range of datasets to determine whether large language models learn implicit representations of

causativity, a linguistic property that is not necessarily represented syntactically or morphologically in English. We compare classifiers based on BERT, DistilBERT, and XLNet, and find that all learn to predict causativity to a large extent. Differences in classification accuracy are observed across different datasets (see Table 1). As expected, all models achieve the highest accuracy on D_{tr_5} and the lowest accuracy on D_{all}. The latter set, in addition to verbs that are lexically causative or lexically non-causative, also includes verbs that participate in the causative-inchoative alternation, which presents an additional challenge to the classifiers.

We also show that causativity is represented rather in the higher layers of the models and, furthermore, that reducing each model to only the 10% of its neurons that are most correlated with the causativity property only leads to small differences in accuracy, sometimes an increase in accuracy due to the elimination of non-discriminative features.

Our error analysis suggests that many of the classification errors are actually labeling errors in the data, due either to a wrong parse of the sentence in our source corpus ENCOW or to the polysemy of verbs that can be causative in certain readings but are not causative in some of the readings mislabeled in the dataset. Put differently, the classifiers were probably better in identifying causativity than their accuracy scores suggest. While our datasets were created with little manual effort and already led to good results, we are planning on pursuing possible improvements in the future in order to avoid these labeling errors as far as possible.

Acknowledgments

The work presented in this paper was partly financed by the Deutsche Forschungsgemeinschaft (DFG) within the project "Unsupervised Frame Induction (FInd)". We wish to thank three anonymous reviewers for their constructive feedback and helpful comments.

References

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, D. Anthony Bau, and James Glass. 2019a. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI, Oral presentation)*.

Fahim Dalvi, Avery Nortonsmith, D. Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. 2019b. Neurox: A toolkit for analyzing individual neurons in neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 9851–9852.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926, Online. Association for Computational Linguistics.

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

David R. Dowty. 1979. *Word Meaning and Montague Grammar*. Springer Netherlands.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Richard Futrell and Roger P. Levy. 2019. Do RNNs learn human-like abstract word order preferences? In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 50–59.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, page 691–696.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4364–4373, Hong Kong, China. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhuoran Liu, Shivali Goel, Mukund Yelahanka Raghuprasad, and Smaranda Muresan. 2019b. Columbia at SemEval-2019 task 7: Multi-task learning for stance classification and rumour verification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1110–1114, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.