Does BERT Recognize an Agent?
Modeling Dowty's Proto-Roles with Contextual Embeddings

Mattia Proietti
CoLing Lab, University of Pisa
m.proietti2@studenti.unipi.it

Gianluca E. Lebani
Ca' Foscari University of Venice
gianluca.lebani@unive.it

Alessandro Lenci
CoLing Lab, University of Pisa
alessandro.lenci@unipi.it

Abstract

Contextual embeddings build multidimensional representations of word tokens based on their context of occurrence. Such models have been shown to achieve a state-of-the-art performance on a wide variety of tasks. Yet, the community struggles in understanding what kind of semantic knowledge these representations encode. We report a series of experiments aimed at investigating to what extent one of such models, BERT, is able to infer the semantic relations that, according to Dowty's Proto-Roles theory, a verbal argument receives by virtue of its role in the event described by the verb.

This hypothesis were put to test by learning a linear mapping from the BERT's verb embeddings to an interpretable space of semantic properties built from the linguistic dataset by White et al. (2016). In a first experiment we tested whether the semantic properties inferred from a typed version of the BERT embeddings would be more linguistically plausible than those produced by relying on static embeddings. We then move to evaluate the semantic properties inferred from the contextual embeddings both against those available in the original dataset, as well as by assessing their ability to model the semantic properties possessed by the agent of the verbs participating in the so-called causative alternation.

1 Introduction

In the last two decades, word embeddings have become one of the most widely used tools for the encoding of lexical meaning in computational models of language. Different flavours of such models have been proposed, all of which have in common the idea of representing lexical elements as multidimensional vectors inferred from their context of occurrence (for a review, see Lenci, 2018).

The last wave of word embeddings followed the transformer-based models breakthrough (Vaswani et al., 2017), that resulted in the development of the so-called **contextual embeddings**. These representations are generated by models like BERT (Devlin et al., 2019) or GPTs (Radford and Narasimhan, 2018; Radford et al., 2019) and derive their name by their ability to keep track of the different contexts in which a word occurs, giving different vector representations for the same word appearing with different surrounding neighbours (for review, see Liu et al., 2020; Ethayarajh, 2019). This has been a major improvement over **static embeddings** obtained from models such as LSA (Landauer and Dumais, 1997), GloVe (Pennington et al., 2014) and Word2Vec (Mikolov et al., 2013), allowing this kind of representation to reach state-of-the-art performance in a great variety of Natural Language Processing (NLP) tasks.

Notwithstanding their wide usage, mainly due to their great empirical successes, the community still struggles to understand what kind of information word embeddings are actually able to encode about language structure, and how they do it. The problem has been even sharpened with contextual embeddings, which are considered to be more entangled representations, usually bigger in dimensions than previous versions, and are obtained from deeper neural models, whose inner working is more complex. Due to this fact, to better understand and explain what kind of structure these models are able to represent is becoming more and more desirable and several research lines started to spring out with this purpose. To the present days the encoding of syntactic knowledge in these model has been more studied than their ability to deal with semantic facets of language (Rogers et al., 2020), but the number of studies in that direction, usually carried out by means of probing tasks developed on top of pre-trained architectures (Vulić et al., 2020), is also growing (Chersoni et al., 2021; Ettinger, 2020).

In the present paper we focus on the modelling of the semantic content of what Dowty (Dowty, 1989, 1991) labelled as **Thematic Proto-Roles**, that are clusters of entailment properties that an arguments derives solely by virtue of its role in the event described by a predicate. Following previous work by Lebani and Lenci (2021) we use a linear transformation mapping between embeddings of verbs represented in BERT's space and a set of interpretable vectors derived from the Universal Decompositional Semantics Dataset on Proto-Roles properties (White et al., 2016), in which human ratings about argument properties have been annotated and collected. However, the present work differs from Lebani and Lenci (2021) under several respects: i.) we deal with contextual embeddings, focusing on those yielded by BERT, ii.) we experiment with representations at the token-level, iii.) we successfully apply sPCA as a de-noising technique, iv.) and we qualitatively address the phenomenon of the causative-inchoative alternation, for which the notion of Semantic (Proto-) Role is crucially relevant.

The goal of this paper is twofold: i.) to test whether the BERT contextual embeddings of a verb encode semantic information concerning the Proto-Role properties held by its arguments, and ii.) to test whether this knowledge can be distilled by means of a linear mapping, thus leading the way to the development of full-scale systems able to extract this knowledge for a wide range of verbs.

The following pages are organized as follows: in Section 2, we quickly review the literature investigating the semantic content of vector representations, before discussing the notion of thematic role and its empirical foundations. We describe and test our method in Sections 4 and 5, respectively, while Section 6 is devoted to a general discussion of the merits and limitations of the use of the BERT's embeddings to model Proto-Role information.

2 The Semantics of Word Embeddings

Notwithstanding the wide usage of word embeddings in NLP and related fields, the literature trying to characterize the semantic properties of these representation is quite scarce. Concerning the efforts in trying to understand whether and to what extent thematic roles information is encoded in contextual embeddings, Tenney et al. (2019b), (building on works by Teichert et al., 2017 and Rudinger et al., 2018) proposed a suite of classification tasks aimed at investigating how these representation encode sentence structure across a range of syntactic, semantic, local, and long-range phenomena. Crucially, these authors report a very small improvement over non-contextual baselines. Thematic role information seems to be recoverable with this strategy, but to an extent which is not that notable.

Ettinger (2020) tested BERT on a suite of diagnostics drawn from human language experiments, among which the most relevant to our scopes is the semantic role sensitivity and event knowledge task, that tests the model ability to discern between good and bad sentence fillers on the basis on the required semantic role. Results showed that the model is not that accurate at matching human predictions, even if some of the information appear to be encoded. Klafka and Ettinger (2020) developed a suite of probing tasks with the aim of assessing what kind of semantic information about the surrounding words is encoded in a contextual embedding. For instance, a task of this suite can implement the question "*What does the embedding of the verb tell us about the animacy of the subject noun?*" as a binary task for a MLP classifier trained and tested on the embeddings of a single word (Klafka and Ettinger, 2020, p. 4802). Relevant for our purposes, these scholars report that much information about subject's animacy can be recovered by inspecting the embedding of the verb.

However, our work is more strongly related, both in goals and methods, to those by Fagarasan et al. (2015), Utsumi (2020), Chersoni et al. (2021) and above all to Lebani and Lenci (2021). All these authors have used a linear transformation to learn a mapping between an embedding space and a space derived from human judgements. Fagarasan et al. (2015) learned a mapping towards the short normalized descriptions (*feature norm*) collected by McRae et al. (2005) in order to learn to predict perceptual features for novel words. Both Utsumi (2020) and Chersoni et al. (2021) applied that strategy to decode word vectors in terms of the brain based semantic features collected in Binder et al. (2016). Finally, Lebani and Lenci (2021) focused on finding fine-grained Proto-Role information by learning the mapping between several static embeddings and an entailment space based on the same ratings we use in this experiment, i.e. the White et al. (2016) dataset that is introduced in the next section. All in all, these works show that both contextual and static embeddings encode a wide range of (psycho)linguistic relevant information that can be inferred by means of a simple linear mapping.

3 Thematic Proto-Roles

Theories about the role played by the arguments of the verb at the syntax-semantics interface have come in many flavours. One of the most debated and controversial matter in theoretical linguistics concerns the definition of the notion of **semantic role** and the development of a reliable method for the identification of such categories.

In contrast with the traditional view of semantic roles as discrete primitive semantic categories, David Dowty proposed to reduce the total number of roles to two prototypical notions, which he called **Proto-Agent** and **Proto-Patient** (Dowty, 1989, 1991). In Dowty's view, roles are defined by a set of entailments determined by the meaning of the verb. Some properties contribute to characterize an argument as *Proto-Agent*, while others as *Proto-Patient*, but they can be present in different degrees and in mixed configurations. These configurations correspond to classical intermediate roles such as experiencer, theme, and so on.

This approach has some important advantages over classical views of semantic roles, which have always been in tension between the choice of the right number of roles and the right mapping between grammatical and semantic role. Dowty's theory does not discard completely the possibility of identifying a different role for each different peculiar argument, placing a distinction between *specific roles* and *linguistic roles*. While the former are specific for each verb (e.g., to build has two main arguments: the *builder* and the *buildee*), the latter are generalizations aimed at capturing common traits about different specific roles. For example *builder, killer, worker, seller* are all instances of *Proto-Agent*, but all at a different degree.

Furthermore, concerning the selection of the argument, Dowty sets a straightforward rule in his **Argument Selection Principle** stating that: "*In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of* Proto-Agent *properties will be lexicalized as subject of the predicate; the argument having the greatest number of* Proto-Patient *entailments will be lexicalized as the direct object.*" (Dowty, 1991, p. 576). This claim received empirical validation on a cognitive perspective by Kako (2006), among others. This scholar proved, through a series of experiments, not only that the hypothesis has psychological validity, but also that "*speakers can make inferences about these properties from grammatical roles alone [...]*" (Kako, 2006, p. 34). Inspired by these findings Reisinger et al. (2015) built a crowd-sourcing experiment to test the theory against a large amount of data and substantially confirmed the results by Kako (2006). The latter approach has been the precursor of the dataset by White et al. (2016) that we adopt here to build a semantic space based on human judgments.

3.1 Human Judgements about Proto-Roles

As will be explained in more detail in Section 4, in order to infer whether the BERT contextual embeddings are able to encode some information about semantic roles, we studied the output of a mapping from the contextual embeddings of a group of selected verbs and the ratings produced by a group of speakers. In our experiment we rely on the judgments collected by White et al. (2016). This dataset was built by asking a group of native speakers to read a series of sentences with a highlighted argument and to answer, on a five points Likert scale, to a group of Dowty-inspired questions on the target argument. For example, to know how plausible is for an argument to have a property like `awareness`, the subjects were asked: "*ARG was/were aware of being involved in PRED?*".

The paradigm used by White et al. (2016) was developed from that described by Reisinger et al. (2015). In the latter work the authors annotated sentences from PropBank (Palmer et al., 2005) while White et al. (2016) used the English Web Treebank (Silveira et al., 2014), which is annotated following the Universal Dependencies guidelines (de Marneffe et al., 2021) and covers a greater variety of genres. Furthermore, White et al. (2016) revised the inventory of questions and the method to present them and used redundant annotations. The semantic decomposition principle behind the whole paradigm is well suited to Dowty's theory of Proto-Roles, and vice-versa, due to their common target of reducing semantics categories to smaller dimensions of meaning. This reduction allows not only linguists to better describe the categories, but also naive speakers to understand the questions to characterize the semantic roles.

	⟨nsubj, awareness⟩	⟨nsubj, change of state⟩	⟨nsubj, volition⟩	⟨dobj, awareness⟩	⟨dobj, change of state⟩	⟨dobj, volition⟩
`to affect`	0	0.625	0	0.688	0.75	0.187
`to amaze`	0.25	0.25	0.25	1	0.708	0.792
`to bring`	0.922	0.422	0.828	0.562	0.472	0.319
`to fill`	0.875	0.25	0.875	0.5	0.875	0.562
`to give`	0.899	0.352	0.887	0.062	0.312	0.081
`to ignore`	1	0.875	1	0.75	0.5	0.125
`to include`	0.458	0.51	0.433	0.451	0.461	0.446
`to kill`	0.925	0.65	0.875	0.575	0.937	0.042
`to put`	0.833	0.492	0.84	0.275	0.75	0.11
`to tell`	0.99	0.357	0.959	0.968	0.561	0.714

Table 1: Portion of the entailment-based vector space (adapted from Lebani and Lenci (2021)).

4 General Methodology

Our ultimate goal is to probe the kind of distributional knowledge encoded in contextual embeddings in order to assess whether BERT (and arguable other models of the same family) is able to encode Proto-Role semantic information. As such, we opted for a methodology that has been tested and proven in our reference literature (Fagarasan et al., 2015; Utsumi, 2020; Chersoni et al., 2021; Lebani and Lenci, 2021). Similarly to what has been done by Lebani and Lenci (2021), indeed, we created a linear mapping between a semantic space composed of BERT embeddings and a vector space derived from the ratings collected in the White et al. (2016)'s Proto-Roles dataset.

Model We tested BERT (Devlin et al., 2019) in its `bert-large-cased` version as released in the Hugging Face python library (Wolf et al., 2019). This deep encoder architecture has 24 layers, 1,024 hidden units per layer, 16 attention heads and a total of 336M of parameters. It is pre-trained with masked language modeling and next sentence prediction tasks. As we want to know the semantic properties that BERT encodes in its native representations, we did not fine-tune the model.

Corpus The sentences annotated by White et al. (2016) were extracted from the English Web Treebank (Silveira et al., 2014) corpus, which is available in the Universal Dependencies repository.[1] From the training set of this corpus we extracted a list of 2226 pre-tokenized sentences that were later processed with BERT. From these sentences we extracted only the verb embeddings either at type or token level. For the type-level experiment, verb vectors were averaged across different contexts (Bommasani et al., 2020)[2].

Ratings-based semantic spaces We built different rating-based semantic spaces for the type-level and for the token-level analyses. For the type-level analysis we followed Lebani and Lenci (2021) and built a unique semantic space for both arguments, as shown in Table 1. For the evaluation of the token-level embeddings, on the other side, we built different spaces for the `nsubj` and for `dobj` syntactic roles, choosing to ignore the passive subjects in order to remove excessive sparsity. The latter procedure left us with 1972 token instances for the `nsubj` space and 797 token instances for the `dobj` space. The dimensions of these spaces correspond to the 14 properties tested by the authors, as ranked by annotators for each token. We indexed each token with the id of its sentence, in order to retrieve it and compare different occurrences of the same verb type.

Learning Algorithm As a mapping strategy, in the wake of previous works (Chersoni et al., 2021; Fagarasan et al., 2015; Lebani and Lenci, 2021)

[1] `https://universaldependencies.org/`

[2] We did not need to average between word pieces, as also suggested in Bommasani et al. (2020), since we used sentences that were pre-tokenized at the word level.

we used the Partial Least Squares (PLS) regression implementation in the Scikit-learn Python library (Pedregosa et al., 2011), with the number of components set to 10 and within a ten-fold cross-validation. We evaluated the predicted vectors by calculating its Spearman's rank correlation with the original ones, both row-wise and column-wise.

To check the quality of our model, we generated a matrix for each experiment with values randomly sampled from the interval $[0, 1]$, shaping it like the corresponding BERT space dimensions. We treated the performance of the mapping learned from these randomly generated spaces as a baseline.

5 Experiments

5.1 Experiment 1: Type-level

For the *type-level* analysis, we reproduced on the BERT vectors the experimental settings as in Lebani and Lenci (2021), mainly in order to obtain a set of comparable results. The application of the same filtering strategy resulted in a vector space composed of 155 rows, one for each verb lemma, and 41 columns, corresponding to features made of <*grammatical_function,property*> pairs, which were first aggregated by averaging between different annotators judgments and instances of the same lemma and than scaled to fit the range [0,1].

After constructing the BERT type embeddings as described in Section 4, we moved on to learning the mapping and, as we wanted to have a grasp of the differences in performance across the whole BERT model, we tested each of the layers averaging between them in groups of four (e.g., layers 1-4, layers 5-8, etc). Even if we weren't able to identify significant differences across groups, we found a peak around the group of layers 13-16. We believe that these results are consistent with the findings in Tenney et al. (2019a), where it is shown that syntactic and grammatical information (e.g., word order, POS) is better encoded at lower layers, while semantics features (e.g., semantic roles, coreference) are better represented at higher layers, although the latter seems to be more equally distributed than the former across the whole model. Correlation results for our best performing group of layers – comprising layers 13-16 – are reported below in Table 2, while scores for other groups can be found in Table 3 in Appendix A. We obtained average values directly comparable to the best performing model found in Lebani and Lenci (2021), which is a Skip-Gram model with negative sampling and syntactic typing (SGNS.syn). All the groups of layers we tested performed equally or slightly better in absolute terms than the SGNS.syn used in Lebani and Lenci (2021). In particular, our best performing group of layers reaches average correlation values of $\rho = 0.74$ row-wise (i.e., correlations by verb) and $\rho = 0.39$ column-wise (i.e., correlations by property), against the respective values of SGNS.syn of $\rho = 0.71$ and $\rho = 0.31$. However, we encountered the same problem of Lebani and Lenci (2021) when evaluating the mapping row-wise, that is an unexpected high baseline, which in our case set itself at $\rho = 0.64$[3]. We tried to overcome this limit in the second experiment made at the *token-level*. Overall, the interpretation suggested by those results seems to be that, when reduced to static embeddings, BERT contextual vectors perform only slightly better, if at all, than those of a classic non-contextual Distributional Semantic Model with proper hyper-parameters settings, when it comes to retrieve fine-grained information about thematic Proto-Roles.

Model	Row-wise	Column-wise
BERT$_{13\_16}$	0.74	0.39
baseline_BERT	0.64	-0.07
SGNS.syn	0.71	0.31
baseline_SGNS	0.62	0.035

Table 2: Average Spearman's scores for BERT, group of layers 13-16, and SGNS.syn (the best performing static model in Lebani and Lenci, 2021), with relative baselines.

Regarding correlations obtained by property at the type-level, which are shown in detail in Figure 1, we found that the `dobj` argument seems to be the one easier to model but, interestingly enough, it shows higher values for *Proto-Agent* properties, reaching the highest in `awareness` with a $\rho = 0.62$. There are three properties specifically related to the *Proto-Patient* that are scored relatively high for the `dobj` argument: `change of possession` ($\rho = 0.59$), `change of state` ($\rho = 0.52$) and `was used` ($\rho = 0.49$). As for the `nsubj` argument, the best modeled properties seem to be those that characterize a *Proto-Agent* role, that is to say `sentient` ($\rho = 0.53$), `volition` ($\rho = 0.49$), `awareness` ($\rho = 0.47$)

[3] The baseline scores attested in our trials belonged to the $[0.62, 0.66]$ interval, coherently with the baseline score $\rho = 0.62$ reported by Lebani and Lenci (2021).

and was for benefit (ρ = 0.43). Overall, these results are consistent with those found in Lebani and Lenci (2021) but partially deviate from those we obtained at the token-level, as it will be shown in the next sections.

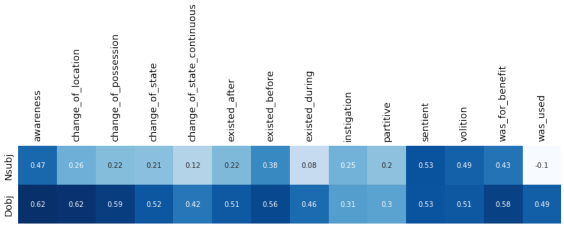

Figure 1: Detailed properties correlation values at the type-level for nsubj and dobj.

5.2 Experiment 2: Token-level

As already mentioned, we created two different mappings at the token-level, one for nsubj argument and one for the dobj argument. Differently from the type-level experiment, in which every data point was an abstract representation of a decontextualised verb lemma described by the 41 features made of the union of grammatical function and Proto-Role property, here we deal with words tokens in context. Since not every verb occurrence received annotations both for nsubj and dobj in White et al. (2016), we created the two sub-spaces, in which each instance is described by the set of fourteen properties elicited with the questions of the SPR2 protocol found in White et al. (2016). In this case we ran the experiment on single layers chosen among those of the best performing group in the previous experiment, which was that of layers 13-16, and we report here the results for layer 16. Simply reproducing the mapping on those spaces with the same settings as the type-level gave us results really similar to the first experiment.

Concerning the high baseline problem, we took into account the hypothesis put forward by Lebani and Lenci (2021). They considered a possible justification of these results the fact that "*Subjects tend to have proto-agent properties, while object tend to have proto-patient properties. From this association*[...] *follows the fact that the vectors of our target entailment-based space are, to a certain extent, bound to share a similar structure in which some dimensions tend to be consistently scored higher than others.*" (Lebani and Lenci, 2021). This proposal is confirmed by measuring the cosine similarity among the vectors of the entailment-based spaces. In fact, a look at the average cosine similarity in the semantic spaces gave us values of cos = 0.85 and cos = 0.77, respectively for the nsubj and dobj, showing that indeed there is a high similarity score among the vectors, which can introduce noise and alter the learning process of our model, thereby allowing the baseline to reach high correlations. Thus, we tried to use a dimensionality reduction technique such as Sparse Principal Components Analysis (sPCA) in its Scikit-learn implementation, which is based on Mairal et al. (2009). The goal was to introduce sparsity in our data and reduce the noise, without reducing the number of dimensions and losing interpretability. This technique is mostly used for de-noising purposes in the field of computer vision, but rarely employed in NLP (Drikvandi and Lawal, 2020). Differently from classic PCA, sPCA does not yield orthogonal dimensions in the space where it is applied, but seems to succeed in reducing similarity among the instances of our ratings-based spaces. As a matter of fact, the average cosine value, for both nsubj and dobj space, resulted in cos $\approx$ 0 after the application of this technique. Furthermore, the sparsity of the loadings generated with this method allowed us to have a better grasp of which principal component represented which variable.

As shown in Figure 2, we obtained average correlations of $\rho = 0.50$ for nsubj and $\rho = 0.40$ for dobj at the row-level, with the baseline keeping itself at $\rho \approx 0$ in both cases[4] . It should be noted that the reported manipulations with sPCA affect only the row-wise analysis, which was indeed the only one suffering from the high baseline problem. The average values obtained column-wise remain the same, and the same has to be said for the fine grained analysis of single properties. In fact, the correlation values obtained by the new components yielded by sPCA overlap perfectly with the original variables. At the column level we got $\rho = 0.43$ for the nsubj space and $\rho = 0.38$ for the dobj space. Despite the fact that these correlations do not reach outstanding values, they are significantly higher than the baseline both row-wise and column-wise, and the ones obtained with BERT and SGNS type

[4] We also experimented with standard PCA. We searched for a number of components capable of accounting for the 85% of the variance, thus obtaining a different number of components for our spaces: 6 for nsubj and 8 for dobj. The correlations with this strategy are lower than those obtained by using sPCA: $\rho = 0.42$ (by row) and $\rho = 0.33$ (by columns) for the nsubj; of $\rho = 0.38$ (by row) and $\rho = 0.34$ (by column) for the dobj.

vectors.

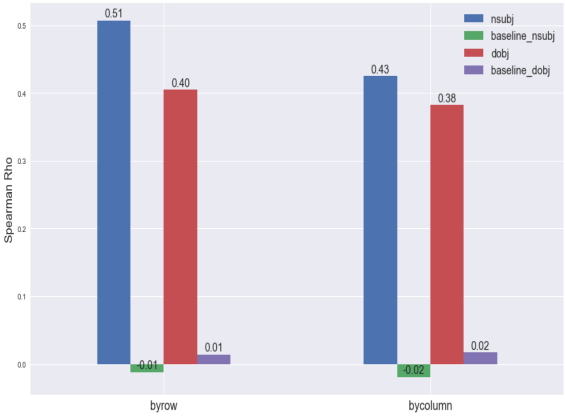

Figure 2: Average correlations at the token-level.

5.3 Experiment 3: Modeling the causative-inchoative alternation

As a third experiment, we decided to make a more qualitative analysis focusing on the so called causative alternation. This linguistic phenomenon is directly tied with theories of semantic/thematic roles and Dowty's theory is no exception. Our hypothesis here is that verbs participating to the causative alternation, occurring both in transitive and intransitive frames, should entail a set of properties more skewed toward the *Proto-Agent* role when appearing in transitive contexts and should incline toward those entailments typical of the *Proto-Patient* in their intransitive occurrences. For example, the verb *to break* can appear in transitive sentences like *John broke the window* or in intransitive ones as *The window broke*, entailing different properties about the respective subjects. In fact, the two subjects are supposed to be realizations of different underlying Proto-Roles, a *Proto-Agent* in the former case, a *Proto-Patient* in the latter.

Our aim was to test BERT embeddings to know whether they are able to encode some information about that alternation. We focused again on the token level, using the nsubj space previously created to train a PLS regression model on transitive verbs. We selected 100 sentences containing 50 alternating verb types, thus having 50 pairs of causative alternation examples. Target verbs for this experiment have been selected following the Levin (1993)'s classification as coded in VerbNet (Schuler, 2006). Sentences containing these target verbs have been extracted manually from a variety of sources, comprising VerbNet frames examples, FrameNet (Baker et al., 1998) examples, and enTenTen corpus through Sketch Engine (Kilgarriff et al., 2014). We found that causative alternation is indeed well modeled in the majority of cases (35 out of 50 pairs of sentences, 70%), as can be seen from Figure 3, which shows a visual representation of a portion of our predicted vector space.

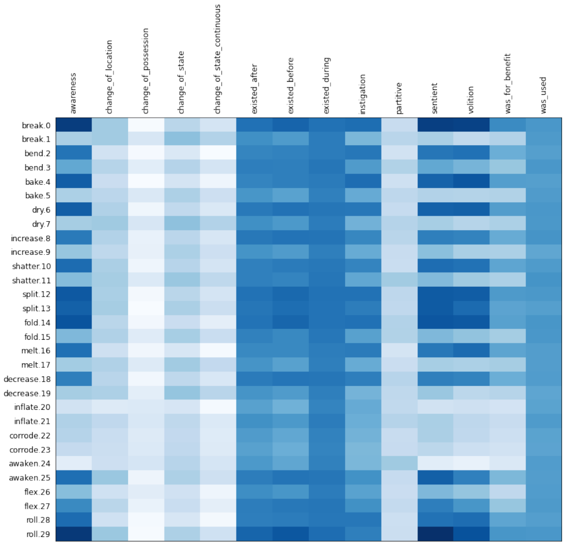

Figure 3: Visualization of the first 30 alternating verbs in our predicted space. Even ids are for transitive frames, odd for intransitive ones.

The alternation is clearly visible in the difference of intensity in those slots of the heatmap corresponding to the *Proto-Agent* properties, mostly in awareness, sentient, volition, instigation and, to a lesser extent, to those corresponding to the *Proto-Patient*, mostly in change of state, change of state continuous, was used. It should be noted that our model fails to catch the alternation in some pairs of verbs (*awaken, flex* and *roll* in the portion showed in Figure 3)[5]. However, these failures, which represent the 30% of the predicted outcomes (15/35), pave the way to further considerations that we discuss in the next section.

6 General Discussion

6.1 Ability to recover fine-grained Proto-Role information in BERT's embeddings

All the three reported experiments show that it is indeed possible to recover Proto-Role information about the arguments from verb embeddings, as demonstrated by the average correlation values

[5] Due to an error, ids for the verb *inflate* are switched. Thus it seems that the intransitive has higher scores in *Proto-Agent* properties than the transitive, which is the contrary.

obtained both row-wise and column-wise, which are significantly higher than those of the baseline, mostly in the token-level experiment in which a preliminary sPCA transformation has been applied. An even more fine-grained analysis can be conducted taking into account single properties. Although not directly comparable, the results at the token-level partially contradict the trend obtained at the type-level, such that in the former the `nsubj` argument shows higher correlation values than the `dobj`, while for the latter the contrary happens. However, what they have in common is the fact that in general *Proto-Agent* properties seem to be better modeled in both experiments and for both arguments. As a matter of fact, at the token-level this happen not only, for the `nsubj`, as expected, but also for the `dobj` even if, in this second experiment, we obtained higher correlations for the `nsubj` argument, as can be seen in Figure 4. In particular, three properties seem to be pretty well modeled: `awareness, sentient` and `volition`, which are the core entailments of the *Proto-Agent* role, and are strongly related to the animacy of typical subject arguments.

On the contrary, our model struggles to cope with *Proto-Patient* properties at the token-level in both spaces. Whether this evidence means that *Proto-Agent* properties are better represented in BERT or in just the rating-based space it is not easy to say. But, from a theoretical point of view it should be considered that the individuation of good examples of properties for the *Proto-Patient* has been an issue ever since the statement of Dowty's theory. In fact, Dowty himself claimed that "*Proto-Patient properties are harder to isolate entirely*" (Dowty, 1991, p. 576) than those of the *Proto-Agent*.

Moreover, both Reisinger et al. (2015) and Kako (2006) found out in their experiments that "*Proto-Agent properties have a greater effect than Proto-Patient properties*"(Reisinger et al., 2015, p. 481). All these cues might suggest that the individuation and the modeling of *Proto-Patient* properties might be a more difficult matter than *Proto-Agent* ones and that the latter have a more solid stand from several point of views: theoretical analysis, cognitive and corpus-level testing, and probably even in the knowledge encoding operated by BERT. Also, it should be taken into consideration that some further developments of Dowty's theory dispensed with *Proto-Patient* properties at all, building only on those of the *Proto-Agent* and characterizing its opposite role in negative terms (see, for example, the theory elaborated by Grimm, 2011).

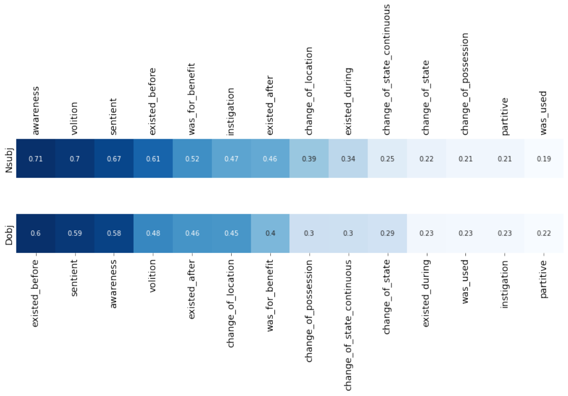

Figure 4: Detailed single properties correlations for `nsubj` and `dobj`.

6.2 Modeling the causative-inchoative alternation

As it has been shown, we have been able through our strategy to model the causative-inchoative alternation in terms of Proto-Role properties prediction. In fact in 70% of the pairs we predicted, the transitive version of the verb scored higher values in *Proto-Agent* properties. In particular, those cases representing prototypical instances of such phenomenon are almost perfectly predicted. Consider as an example, the pair of sentences regarding the verb *to break* (*break.0* and *break.1* in Figure 3). The corresponding sentences are taken directly from VerbNet example frames and are: *Tony broke the window* and *The window broke*, respectively for the transitive and the intransitive frame. The properties seem to be well predicted not only for those concerning proto-agency, but also for those entailments of the *Proto-Patient*. That is to say, while the first verb, *break.0*, which is transitive, shows a greater intensity (i.e., higher predicted values) than *break.1* in the slots corresponding to `awareness, instigation, volition, sentient, existed after, existed before` and `was for benefit`, the reverse is true if we consider *Proto-Patient* entailments. In fact, the intransitive *break.1* has greater values in `change of possession, change of state, change of state continuous, partitive`. This is a recurrent pattern among all the predicted space. Moreover, Figure 5 shows how on average *Proto-Agent* properties are scored higher in the predicted subspace formed by only the transitive use of the verbs. Vice-versa almost all the *Proto-Patient* properties have higher mean scores among the intransitive use, even if with a much smaller difference. This is consistent with the assumption that subjects of inaccusative verbs are less agentive than those of their transitive counterpart. However, there are a few notable exceptions to this trend. Three properties in particular seem to contradict the assumptions of the theory: `change of location`, `partitive` and `was used`. The first is assumed to be typical of *Proto-Agent* and, instead, shows a higher average score for the transitive use. The second and the third, supposed to be *Proto-Patient* properties, reach the same values in both sub-spaces. Among the 15/50 pairs (30%) which our model failed to predict, there are 6 instances in which the subject of the intransitive verb is more animate than that of the transitive verb. We regard this fact as a possible influence in the prediction, due to the fact that animacy and agency, and consequently their characterizing properties, are two strictly related concepts and they might even overlap in some circumstances, for example in the determination of subjecthood. Given the contextual nature of the BERT embeddings, it is no surprise that verb representations are adjusted in relation to the other elements of the sentence, incorporating at each occurrence particular information about surrounding words. In particular, the fact that animacy information about the subject is projected into the verb embedding and is recoverable from it has been shown by Klafka and Ettinger (2020).

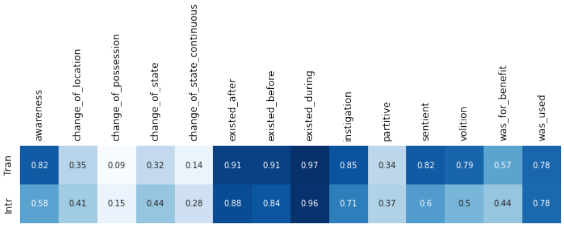

Figure 5: Average predicted properties for transitive and intransitive use.

7 Conclusions

Although our strategy has been proven to be good at modeling the Proto-Roles phenomenon in BERT embeddings to a certain extent, some intrinsic limitations of our work have to be taken into account. Firstly, we used a linear regression (PLS) model as a strategy to build the mapping, but, due to the complexity of the type of information enquired and that of the BERT space, more complex, non-linear transformations, like a Multi Layer Perceptron, might be a better choice for the task. Secondly, the data we used are the best at our disposal, but they are not necessarily the best possible in absolute and might be further improved, by both revising the questions and the set of properties. Thirdly, we obtained mixed results between the token and the type levels concerning which is the best modeled Proto-Role.

Notwithstanding these limitations, we have shown that fine-grained information about Proto-Roles properties of the arguments is recoverable inspecting the embeddings of the verbs yielded by BERT. Also, our results suggest that there might be a discrepancy between the properties of the two Proto-Roles and that *Proto-Agent* properties are better modeled and predicted. We have also been able to show how different Proto-Roles entailments can be predicted in verbs participating to the causative-inchoative alternation. Additionally, we successfully employed sPCA to reduce the noise in our data, which might be a promising cue about future usages of this technique in the field of NLP.

It is worth emphasizing that the main goal of this research is to test BERT's ability to capture some crucial aspects of the verbal argument structure. Even if there can be practical applications of our method (e.g., it can be used as a starting point for Semantic Role labeling or, crucially, Semantic Proto-role labeling; Reisinger et al. 2015; Teichert et al. 2017), our main interest is more theoretical and methodological. Many of the probing tasks that are used today, indeed, do not focus on the prototypical nature of semantic roles, which is precisely a fundamental pillar and the major innovation of Dowty's theory and of the present work. Moreover, our analysis of the causative/inchoative alternation is just a first example of a series of tasks that we plan to develop to characterize the knowledge acquired by these models to explore key aspects of the syntax-semantic interface and of verb argument realization. Finally, we will also extend our approach to other contextual embeddings models, like GPT (Radford and Narasimhan, 2018; Radford et al., 2019) and XLNet (Yang et al., 2019).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Jeffrey R. Binder, Lisa L. Conant, Colin J. Humphries, Leonardo Fernandino, Stephen B. Simons, Mario Aguilar, and Rutvik H. Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33:130 – 174.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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. 1989. On the semantic content of the notion of 'thematic role'. In Gennaro Chierchia, Barbara H. Partee, and Raymond Turner, editors, *Properties, Types and Meaning: Volume II: Semantic Issues*, pages 69–129. Springer Netherlands, Dordrecht.

David R. Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67:547 – 619.

Reza Drikvandi and Olamide O. Lawal. 2020. Sparse principal component analysis for natural language processing. *Annals of Data Science*, pages 1–17.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

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

Luana Fagarasan, Eva Maria Vecchi, and Stephen Clark. 2015. From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 52–57, London, UK. Association for Computational Linguistics.

Scott Grimm. 2011. Semantics of case. *Morphology*, 21:515–544.

Edward Kako. 2006. Thematic role properties of subjects and objects. *Cognition*, 101:1–42.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, pages 7–36.

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, Online. Association for Computational Linguistics.

Thomas K. Landauer and Susan T. Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

Gianluca E. Lebani and Alessandro Lenci. 2021. Investigating dowty's proto-roles with embeddings. *Lingue e Linguaggio*, 20:165–197.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4(1):151–171.

Beth Levin. 1993. *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Qi Liu, Matt J. Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *ArXiv*, abs/2003.07278.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 689–696, New York, NY, USA. Association for Computing Machinery.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37:547–559.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

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, Doha, Qatar. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pretraining. In *arxiv*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

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.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. 2018. Neural-Davidsonian semantic proto-role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 944–955, Brussels, Belgium. Association for Computational Linguistics.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Adam Teichert, Adam Poliak, Benjamin Van Durme, and Matthew Gormley. 2017. Semantic Proto-Role Labeling. In *AAAI Conference on Artificial Intelligence (AAAI)*.

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. *ArXiv*, abs/1905.06316.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive science*, 44 6:e12844.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

A Appendix

Here we report the results obtained computing the average Spearman's Rho between predicted vectors and original ones at the type-level, both by row and by column. We indicate each group with the formula BERT$_{x\_y}$ in which x is the first layer of the group and y is the last one.

Model	By Row	By column
BERT$_{1\_4}$	0.71	0.31
BERT$_{5\_8}$	0.73	0.36
BERT$_{9\_12}$	0.72	0.34
BERT$_{13\_16}$	**0.74**	**0.39**
BERT$_{17\_20}$	0.73	0.36
BERT$_{21\_24}$	0.72	0.34
baseline_BERT	0.64	-0.07
SGNS.syn	0.71	0.31
baseline_SGNS	0.62	0.035

Table 3: Average correlation values obtained for each group of BERT layers and for the best performing model in Lebani and Lenci (2021), a SGNS.syn, with relative baselines. Analysis at the type-level.