A Matter of Framing:
The Impact of Linguistic Formalism on Probing Results

Ilia Kuznetsov and **Iryna Gurevych**

Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
`http://www.ukp.tu-darmstadt.de/`

Abstract

Deep pre-trained contextualized encoders like BERT (Devlin et al., 2019) demonstrate remarkable performance on a range of downstream tasks. A recent line of research in probing investigates the linguistic knowledge implicitly learned by these models during pre-training. While most work in probing operates on the task level, linguistic tasks are rarely uniform and can be represented in a variety of formalisms. Any linguistics-based probing study thereby inevitably commits to the formalism used to annotate the underlying data. Can the choice of formalism affect probing results? To investigate, we conduct an in-depth cross-formalism layer probing study in role semantics. We find linguistically meaningful differences in the encoding of semantic role- and proto-role information by BERT depending on the formalism and demonstrate that layer probing can detect subtle differences between the implementations of the same linguistic formalism. Our results suggest that linguistic formalism is an important dimension in probing studies and should be investigated along with the commonly used cross-task and cross-lingual experimental settings.

1 Introduction

The emergence of deep pre-trained contextualized encoders has had a major impact on the field of natural language processing. Boosted by the availability of general-purpose frameworks like AllenNLP (Gardner et al., 2018) and Transformers (Wolf et al., 2019), pre-trained models like ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019) have caused a shift towards simple architectures where a strong pre-trained encoder is paired with a shallow downstream model, often outperforming the intricate task-specific architectures of the past.

The versatility of pre-trained representations implies that they encode some aspects of general
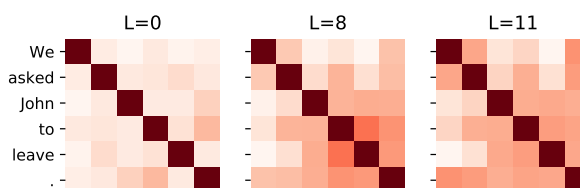

Figure 1: Intra-sentence similarity by layer L of the multilingual BERT-base. Functional tokens are similar in $L = 0$, syntactic groups emerge at higher levels.

linguistic knowledge (Reif et al., 2019). Indeed, even an informal inspection of layer-wise intra-sentence similarities (Fig. 1) suggests that these models capture elements of linguistic structure, and those differ depending on the layer of the model. A grounded investigation of these regularities allows to interpret the model's behaviour, design better pre-trained encoders and inform the downstream model development. Such investigation is the main subject of probing, and recent studies confirm that BERT implicitly captures many aspects of language use, lexical semantics and grammar (Rogers et al., 2020).

Most probing studies use linguistics as a theoretical scaffolding and operate on a task level. However, there often exist multiple ways to represent the same linguistic phenomenon: for example, English dependency syntax can be encoded using a variety of *formalisms*, incl. Universal (Schuster and Manning, 2016), Stanford (de Marneffe and Manning, 2008) and CoNLL-2009 dependencies (Hajič et al., 2009), all using different label sets and syntactic head attachment rules. Any probing study inevitably commits to the specific theoretical framework used to produce the underlying data. The differences between linguistic formalisms, however, can be substantial.

Can these differences affect the probing results? This question is intriguing for several reasons. Lin-

171

guistic formalisms are well-documented, and if the choice of formalism indeed has an effect on probing, cross-formalism comparison will yield new insights into the linguistic knowledge obtained by contextualized encoders during pre-training. If, alternatively, the probing results remain stable despite substantial differences between formalisms, this prompts a further scrutiny of what the pre-trained encoders in fact encode. Finally, on the reverse side, cross-formalism probing might be used as a tool to empirically compare the formalisms and their language-specific implementations. To the best of our knowledge we are the first to explicitly address the influence of formalism on probing.

Ideally, the task chosen for a cross-formalism study should be encoded in multiple formalisms using the same textual data to rule out the influence of the domain and text type. While many linguistic corpora contain several layers of linguistic information, having the same textual data annotated with multiple formalisms for the *same* task is rare. We focus on role semantics – a family of shallow semantic formalisms at the interface between syntax and propositional semantics that assign roles to the participants of natural language utterances, determining *who* did *what* to *whom*, *where*, *when* etc. Decades of research in theoretical linguistics have produced a range of role-semantic frameworks that have been operationalized in NLP: syntax-driven PropBank (Palmer et al., 2005), coarse-grained VerbNet (Kipper-Schuler, 2005), fine-grained FrameNet (Baker et al., 1998), and, recently, decompositional Semantic Proto-Roles (SPR) (Reisinger et al., 2015; White et al., 2016). The SemLink project (Bonial et al., 2013) offers parallel annotation for PropBank, VerbNet and FrameNet for English. This allows us to isolate the object of our study: apart from the role-semantic labels, the underlying data and conditions for the three formalisms are identical. SR3DE (Mújdricza-Maydt et al., 2016) provides compatible annotation in three formalisms for German, enabling cross-lingual validation of our results. Combined, these factors make role semantics an ideal target for our cross-formalism probing study.

A solid body of evidence suggests that encoders like BERT capture syntactic and lexical-semantic properties, but only few studies have considered probing for predicate-level semantics (Tenney et al., 2019b; Kovaleva et al., 2019). To the best of our knowledge we are the first to conduct a cross-formalism probing study on role semantics, thereby contributing to the line of research on how and whether pre-trained BERT encodes higher-level semantic phenomena.

Contributions. This work studies the effect of the linguistic formalism on probing results. We conduct cross-formalism experiments on PropBank, VerbNet and FrameNet role prediction in English and German, and show that the **formalism can affect probing results** in a linguistically meaningful way; in addition, we demonstrate that layer probing can detect subtle differences between implementations of the same formalism in different languages. On the technical side, we advance the recently introduced edge and layer probing framework (Tenney et al., 2019b); in particular, we introduce **anchor tasks** - an analytical tool inspired by feature-based systems that allows deeper qualitative insights into the pre-trained models' behaviour. Finally, advancing the current knowledge about the encoding of predicate semantics in BERT, we perform a fine-grained semantic proto-role probing study and demonstrate that **semantic proto-role properties can be extracted from pre-trained BERT**, contrary to the existing reports. Our results suggest that along with task and language, linguistic formalism is an important dimension to be accounted for in probing research.

2 Related Work

2.1 BERT as Encoder

BERT is a Transformer (Vaswani et al., 2017) encoder pre-trained by jointly optimizing two unsupervised objectives: masked language model and next sentence prediction. It uses WordPiece (WP, Wu et al. (2016)) subword tokens along with positional embeddings as input, and gradually constructs sentence representations by applying token-level self-attention pooling over a stack of layers L. The result of BERT encoding is a layer-wise representation of the input wordpiece tokens with higher layers representing higher-level abstractions over the input sequence. Thanks to the joint pre-training objective, BERT can encode words and sentences in a unified fashion: the encoding of a sentence or a sentence pair is stored in a special token *[CLS]*.

To facilitate multilingual experiments, we use the multilingual BERT-base (mBERT) published by Devlin et al. (2019). Although several recent encoders have outperformed BERT on benchmarks

(Liu et al., 2019; Lan et al., 2019; Raffel et al., 2019), we use the original BERT architecture, since it allows us to inherit the probing methodology and to build upon the related findings.

2.2 Probing

Due to space limitations we omit high-level discussions on benchmarking (Wang et al., 2018) and sentence-level probing (Conneau et al., 2018a), and focus on the recent findings related to the representation of linguistic structure in BERT. Surface-level information generally tends to be represented in the lower layers of deep encoders, while higher layers store hierarchical and semantic information (Belinkov et al., 2017; Lin et al., 2019). Tenney et al. (2019a) show that the abstraction strategy applied by the English pre-trained BERT encoder follows the order of the classical NLP pipeline. Strengthening the claim about linguistic capabilities of BERT, Hewitt and Manning (2019) demonstrate that BERT implicitly learns syntax, and Reif et al. (2019) show that it encodes fine-grained lexical-semantic distinctions. Rogers et al. (2020) provide a comprehensive overview of BERT's properties discovered to date.

While recent results indicate that BERT successfully represents lexical-semantic and grammatical information, the evidence of its high-level semantic capabilities is inconclusive. Tenney et al. (2019a) show that the English PropBank semantics can be extracted from the encoder and follows syntax in the layer structure. However, out of all formalisms PropBank is most closely tied to syntax, and the results on proto-role and relation probing do not follow the same pattern. Kovaleva et al. (2019) identify two attention heads in BERT responsible for FrameNet relations. However, they find that disabling them in a fine-tuning evaluation on the GLUE (Wang et al., 2018) benchmark does not result in decreased performance.

Although we are not aware of any systematic studies dedicated to the effect of formalism on probing results, the evidence of such effects is scattered across the related work: for example, the aforementioned results in Tenney et al. (2019a) show a difference in layer utilization between constituents- and dependency-based syntactic probes and semantic role and proto-role probes. It is not clear whether this effect is due to the differences in the underlying datasets and task architecture or the formalism per se.

Our probing methodology builds upon the edge and layer probing framework. The encoding produced by a frozen BERT model can be seen as a layer-wise snapshot that reflects how the model has constructed the high-level abstractions. Tenney et al. (2019b) introduce the edge probing task design: a simple classifier is tasked with predicting a linguistic property given a pair of spans encoded using a frozen pre-trained model. Tenney et al. (2019a) use edge probing to analyse the layer utilization of a pre-trained BERT model via scalar mixing weights (Peters et al., 2018) learned during training. We revisit this framework in Section 3.

2.3 Role Semantics

We now turn to the object of our investigation: role semantics. For further discussion, consider the following synthetic example:

a. *[John]$_{Ag}$ gave [Mary]$_{Rc}$ a [book]$_{Th}$.*
b. *[Mary]$_{Rc}$ was given a [book]$_{Th}$ by [John]$_{Ag}$.*

Despite surface-level differences, the sentences express the same meaning, suggesting an underlying semantic representation in which these sentences are equivalent. One such representation is offered by role semantics - a shallow predicate-semantic formalism closely related to syntax. In terms of role semantics, *Mary*, *book* and *John* are **semantic arguments** of the **predicate** *give*, and are assigned **roles** from a pre-defined inventory, for example, Agent, Recipient and Theme.

Semantic roles and their properties have received extensive attention in linguistics (Fillmore, 1968; Levin and Rappaport Hovav, 2005; Dowty, 1991) and are considered a universal feature of human language. The size and organization of the role and predicate inventory are subject to debate, giving rise to a variety of role-semantic formalisms.

PropBank assumes a predicate-independent labeling scheme where predicates are distinguished by their sense (get.01), and semantic arguments are labeled with generic numbered core (Arg0-5) and modifier (e.g. AM-TMP) roles. Core roles are not tied to specific definitions, but the effort has been made to keep the role assignments consistent for similar verbs; Arg0 and Arg1 correspond to the Proto-Agent and Proto-Patient roles as per Dowty (1991). The semantic interpretation of core roles depends on the predicate sense.

VerbNet follows a different categorization scheme. Motivated by the regularities in verb behavior, Levin (1993) has introduced the group-

ing of verbs into intersective classes (ILC). This methodology has been adopted by VerbNet: for example, the VerbNet class `get-13.5.1` would include verbs *earn*, *fetch*, *gain* etc. A verb in VerbNet can belong to several classes corresponding to different senses; each class is associated with a set of roles and licensed syntactic transformations. Unlike PropBank, VerbNet uses a set of approx. 30 thematic roles that have universal definitions and are shared among predicates, e.g. `Agent`, `Beneficiary`, `Instrument`.

FrameNet takes a meaning-driven stance on the role encoding by modeling it in terms of frame semantics: predicates are grouped into frames (e.g. `Commerce_buy`), which specify role-like slots to be filled. FrameNet offers fine-grained frame distinctions, and roles in FrameNet are frame-specific, e.g. `Buyer`, `Seller` and `Money`. The resource accompanies each frame with a description of the situation and its core and peripheral participants.

SPR follows the work of Dowty (1991) and discards the notion of categorical semantic roles in favor of feature bundles. Instead of a fixed role label, each argument is assessed via a 11-dimensional cardinal feature set including Proto-Agent and Proto-Patient properties like `volitional`, `sentient`, `destroyed`, etc. The feature-based approach eliminates some of the theoretical issues associated with categorical role inventories and allows for more flexible modeling of role semantics.

Each of the role labeling formalisms offers certain advantages and disadvantages (Giuglea and Moschitti, 2006; Mújdricza-Maydt et al., 2016). While being close to syntax and thereby easier to predict, PropBank doesn't contribute much semantics to the representation. On the opposite side of the spectrum, FrameNet offers rich predicate-semantic representations for verbs and nouns, but suffers from high granularity and coverage gaps (Hartmann et al., 2017). VerbNet takes a middle ground by following grammatical criteria while still encoding coarse-grained semantics, but only focuses on verbs and core (not modifier) roles. SPR avoids the granularity-generalization trade-off of the categorical inventories, but is yet to find its way into practical NLP applications.

3 Probing Methodology

We take the edge probing setup by Tenney et al. (2019b) as our starting point. Edge probing aims to predict a label given a pair of contextualized span or word encodings. More formally, we encode a WP-tokenized sentence $[wp_1, wp_2, ...wp_k]$ with a frozen pre-trained model, producing contextual embeddings $[e_1, e_2, ...e_k]$, each of which is a layered representation over $L = \{l_0, l_1, ...l_m\}$ layers, with encoding at layer l_n for the wordpiece wp_i further denoted as e_i^n. A trainable scalar mix is applied to the layered representation to produce the final encoding given the per-layer mixing weights $\{a^0, a^1..a^m\}$ and a scaling parameter γ:

$$\overline{e}_i = \gamma \sum_{l=0}^{m} softmax(a^l)e_i^l$$

Given the source src and target tgt wordpieces encoded as $\overline{e}_{src}$ and $\overline{e}_{tgt}$, our goal is to predict the label y.

Due to its task-agnostic architecture, edge probing can be applied to a wide variety of unary (by omitting tgt) and binary labeling tasks in a unified manner, facilitating the cross-task comparison. The original setup has several limitations that we address in our implementation.

Regression tasks. The original edge probing setup only considers classification tasks. Many language phenomena - including positional information and semantic proto-roles, are naturally modeled as regression. We extend the architecture by Tenney et al. (2019b) and support both classification and regression: the former achieved via softmax, the latter via direct linear regression to the target value.

Flat model. To decrease the models' own expressive power (Hewitt and Liang, 2019), we keep the number of parameters in our probing model as low as possible. While Tenney et al. (2019b) utilize pooled self-attentional span representations and a projection layer to enable cross-model comparison, we directly feed the wordpiece encoding into the classifier, using the first wordpiece of a word. To further increase the selectivity of the model, we directly project the source and target wordpiece representations into the label space, opposed to the two-layer MLP classifier used in the original setup.

Separate scalar mixes. To enable fine-grained analysis of probing results, we train and analyze separate scalar mixes for source and target wordpieces, motivated by the fact that the classifier might utilize different aspects of their representation for prediction[1]. Indeed, we find that the

---

[1]Tenney et al. (2019b, Appendix C) also use separate pro-

mixing weights learned for source and target word-pieces might show substantial – and linguistically meaningful – variation. Combined with regression-based objective, separating the scalar mixes allows us to scrutinize layer utilization patterns for semantic proto-roles.

Sentence-level probes. Utilizing the BERT-specific sentence representation *[CLS]* allows us to incorporate the sentence-level natural language inference (NLI) probe into our kit.

Anchor tasks. We employ two analytical tools from the original layer probing setup. Mixing weight plotting compares layer utilization among tasks by visually aligning the respective learned weight distributions transformed via a softmax function. Layer center-of-gravity is used as a summary statistic for a task's layer utilization. While the distribution of mixing weights along the layers allows us to estimate the order in which information is processed during encoding, it doesn't allow to directly assess the *similarity* between the layer utilization of the probing tasks.

Tenney et al. (2019a) have demonstrated that the order in which linguistic information is stored in BERT mirrors the traditional NLP pipeline. A prominent property of the NLP pipelines is their use of low-level features to predict downstream phenomena. In the context of layer probing, probing tasks can be seen as end-to-end feature extractors. Following this intuition, we define two groups of probing tasks: *target tasks* – the main tasks under investigation, and *anchor tasks* – a set of related tasks that serve as a basis for qualitative comparison between the targets. The softmax transformation of the scalar mixing weights allows to treat them as probability distributions: the higher the mixing weight of a layer, the more likely the probe is to utilize information from this layer during prediction. We use Kullback-Leibler divergence to compare target tasks (e.g. role labeling in different formalisms) in terms of their similarity to lower-level anchor tasks (e.g. dependency relation and lemma). Note that the notion of anchor task is contextual: the same task can serve as a target and as an anchor, depending on the focus of the study.

---

jections in the background, but do not investigate the differences between the learned projections.

	tok	sent	pred	arg
CoNLL+SL	312.2K	11.3K	13.3K	23.9K
SR3de	62.6K	2.8K	2.9K	5.5K

Table 1: Statistics for CoNLL+SemLink (English) and SR3de (German), only core roles.

4 Setup

4.1 Source data

For German we use the SR3de corpus (Mújdricza-Maydt et al., 2016) that contains parallel PropBank, FrameNet and VerbNet annotations for verbal predicates. For English, SemLink (Bonial et al., 2013) provides mappings from the original PropBank corpus annotations to the corresponding FrameNet and VerbNet senses and semantic roles. We use these mappings to enrich the CoNLL-2009 (Hajič et al., 2009) dependency role labeling data – also based on the original PropBank – with roles in all three formalisms via a semi-automatic token alignment procedure. The resulting corpus is substantially smaller than the original, but still an order of magnitude larger than SR3de (Table 1). Both corpora are richly annotated with linguistic phenomena on word level, including part-of-speech, lemma and syntactic dependencies. The XNLI probe is sourced from the corresponding development split of the XNLI (Conneau et al., 2018b) dataset. The SPR probing tasks are constructed from the original data by Reisinger et al. (2015).

	type	en	de
*token.ix	unary	208.9K	46.9K
ttype [v]	unary	177.2K	34.0K
lex.unit [v]	unary	187.6K	35.7K
pos	unary	312.2K	62.6K
deprel	binary	300.9K	59.8K
role	binary	23.9K	5.5K
*spr	binary	9.7K	-
xnli	unary	2.5K	2.5K

Table 2: Probing task statistics. Tasks marked with [v] use a most frequent label vocabulary. Here and further, tasks marked with * are regression tasks.

4.2 Probing tasks

Our probing kit spans a wide range of probing tasks, from primitive surface-level tasks mostly utilized as anchors later to high-level semantic tasks that

language	en	de
PropBank	5	10
VerbNet	23	29
FrameNet	189	300

Table 3: # of role probe labels by formalism.

task	input	label
token.ix	I [saw] a cat.	→ 2
ttype	I [saw] a cat.	→ saw
lex.unit	I [saw] a cat.	→ see.V
pos	I [saw] a cat.	→ VBD
deprel	[I]$_{tgt}$ [saw]$_{src}$ a cat.	→ SBJ
role.vn	[I]$_{tgt}$ [saw]$_{src}$ a cat.	→ Experiencer
spr.vltn	[I]$_{tgt}$ [saw]$_{src}$ a cat.	→ 2

Table 4: Word-level probing task examples for English. `vltn` corresponds to the `volition` SPR property.

aim to provide a representational upper bound to predicate semantics. We follow the training, test and development splits from the original SR3de, CoNLL-2009 and SPR data. The XNLI task is sourced from the development set and only used for scalar mix analysis. To reduce the number of labels in some of the probing tasks, we collect frequency statistics over the corresponding training sets and only consider up to 250 most frequent labels. Below we define the tasks in order of their complexity, Table 2 provides the probing task statistics, Table 3 compares the categorical role labeling formalisms in terms of granularity, and Table 4 provides examples. We evaluate the classification performance using Accuracy, while regression tasks are scored via R^2.

Token type (`ttype`) predicts the type of a word. This requires contextual processing since a word might consist of several wordpieces;

Token position (`token.ix`) predicts the linear position of a word, cast as a regression task over the first 20 words in the sentence. Again, the task is non-trivial since it requires the words to be assembled from the wordpieces.

Part-of-speech (`pos`) predicts the language-specific part-of-speech tag for the given token.

Lexical unit (`lex.unit`) predicts the lemma and POS of the given word – a common input representation for the entries in lexical resources. We extract coarse POS tags by using the first character of the language-specific POS tag.

Dependency relation (`deprel`) predicts the dependency relation between the parent `src` and dependent `tgt` tokens;

Semantic role (`role.[frm]`) predicts the semantic role given a predicate `src` and an argument `tgt` token in one of the three role labeling formalisms: PropBank `pb`, VerbNet `vn` and FrameNet `fn`. Note that we only probe for the role label, and the model has no access to the verb sense information from the data.

Semantic proto-role (`spr.[prop]`) is a set of eleven regression tasks predicting the values of the proto-role properties as defined in (Reisinger et al., 2015), given a predicate `src` and an argument `tgt`.

XNLI is a sentence-level NLI task directly sourced from the corresponding dataset. Given two sentences, the goal is to determine whether an entailment or a contradiction relationship holds between them. We use NLI to investigate the layer utilization of mBERT for high-level semantic tasks. We extract the sentence pair representation via the *[CLS]* token and treat it as a unary probing task.

5 Results

Our probing framework is implemented using AllenNLP.[2] We train the probes for 20 epochs using the Adam optimizer with default parameters and a batch size of 32. Due to the frozen encoder and flat model architecture, the total runtime of the main experiments is under 8 hours on a single Tesla V100 GPU. In addition to pre-trained mBERT we report baseline performance using a frozen untrained mBERT model obtained by randomizing the encoder weights post-initialization as in Jawahar et al. (2019).

5.1 General Trends

While absolute performance is secondary to our analysis, we report the probing task scores on respective development sets in Table 5. We observe that grammatical tasks score high, while core role labeling lags behind - in line with the findings of Tenney et al. (2019a)[3] We observe lower scores for German role labeling which we attribute to the lack of training data. Surprisingly, as we show below, this doesn't prevent the edge probe from

---

[2]Code available: https://github.com/UKPLab/emnlp2020-formalism-probing

[3]Our results are not directly comparable due to the differences in datasets and formalisms.

176

task	en	de
*token.ix	0.95 (0.93)	0.92 (0.87)
ttype	1.00 (0.92)	1.00 (0.48)
lex.unit	1.00 (0.75)	1.00 (0.33)
pos	0.97 (0.40)	0.97 (0.26)
deprel	0.95 (0.42)	0.95 (0.41)
role.fn	0.92 (0.18)	0.59 (0.10)
role.pb	0.96 (0.67)	0.71 (0.49)
role.vn	0.94 (0.47)	0.73 (0.30)

Table 5: Best dev score for word-level tasks over 20 epochs, Acc for classification, R^2 for regression; Baseline in parentheses.

learning to locate relevant role-semantic information in mBERT's layers.

The untrained mBERT baseline expectedly underperforms; however, we note good baseline results on surface-level tasks for English, which we attribute to memorizing token identity and position: although the weights are set randomly, the frozen encoder still associates each wordpiece input with a fixed random vector. We have confirmed this assumption by scalar mix analysis of the untrained mBERT baseline: in our experiments the baseline probes for both English and German attended almost exclusively to the first few layers of the encoder, independent of the task. For brevity, here and further we do not examine baseline mixing weights and only report the scores.

Our main probing results mirror the findings of Tenney et al. (2019a) about the sequential processing order in BERT. We observe that the layer utilization among tasks (Fig. 2) generally aligns for English and German[4], although we note that in terms of center-of-gravity mBERT tends to utilize deeper layers for German probes. Basic word-level tasks are indeed processed early by the model, and XNLI probes focus on deeper levels, suggesting that the representation of higher-level semantic phenomena follows the encoding of syntax and predicate semantics.

5.2 The Effect of Formalism

Using separate scalar mixes for source and target tokens allows us to explore the cross-formalism encoding of role semantics by mBERT in detail. For both English and German role labeling, the probe's layer utilization drastically differs for predicate and

---
[4]Echoing the recent findings on mBERT's multilingual capacity (Pires et al., 2019; Kondratyuk and Straka, 2019).
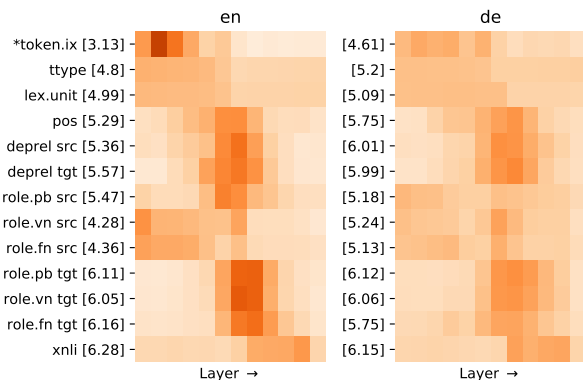

Figure 2: Layer probing results

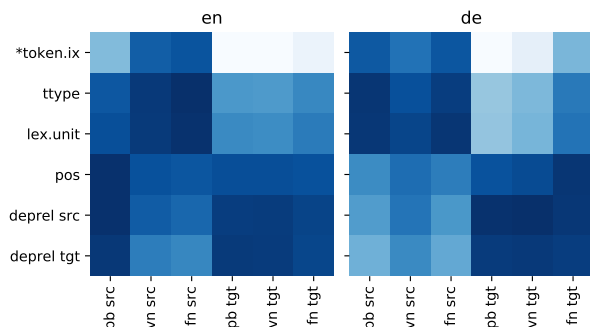Figure 3: Anchor task analysis of SRL formalisms.

argument tokens. While the argument representation role*tgt mostly focuses on the same layers as the dependency parsing probe, the layer utilization of the predicates role*src is affected by the chosen formalism. In English, PropBank predicate token mixing weights emphasize the same layers as dependency parsing – in line with the previously published results. However, the probes for VerbNet and FrameNet predicates (role.vn src and role.fn src) utilize the layers associated with ttype and lex.unit that contain lexical information. Coupled with the fact that both VerbNet and FrameNet assign semantic roles based on lexical-semantic predicate groupings (frames in FrameNet and verb classes in VerbNet), this suggests that the lower layers of mBERT implicitly encode predicate sense information; moreover, sense encoding for VerbNet utilizes deeper layers of the model associated with syntax, in line with VerbNet's predicate classification strategy. This finding confirms that the formalism can indeed have linguistically meaningful effects on probing results.

5.3 Anchor Tasks in the Pipeline

We now use the scalar mixes of the role labeling probes as target tasks, and lower-level probes

177

as anchor tasks to qualitatively explore the differences between how our role probes learn to represent predicates and semantic arguments[5] (Fig. 3). The results reveal a distinctive pattern that confirms our previous observations: while Verb-Net and FrameNet predicate layer utilization `src` is similar to the scalar mixes learned for `ttype` and `lex.unit`, the learned argument representations `tgt` and the PropBank predicate attend to the layers associated with dependency relation and POS probes. Aside from the PropBank predicate encoding which we address below, the pattern reproduces for English and German. This aligns with the traditional separation of the semantic role labeling task into predicate disambiguation followed by semantic argument identification and labeling, along with the feature sets employed for these tasks (Björkelund et al., 2009). Note that the observation about the pipeline-like task processing within the BERT encoders thereby holds, albeit on a sub-task level.

5.4 Formalism Implementations

Both layer and anchor task analysis reveal a prominent discrepancy between English and German role probing results: while the PropBank predicate layer utilization for English mostly relies on syntactic information, German PropBank predicates behave similarly to VerbNet and FrameNet. The lack of systematic cross-lingual differences between layer utilization for other probing tasks[6] allows us to rule out the effect of purely typological features such as word order and case marking as a likely cause.

The difference in the number of role labels for English and German PropBank, however, points at possible qualitative differences in the labeling schemes (Table 3). The data for English stems from the token-level alignment in SemLink that maps the original PropBank roles to VerbNet and FrameNet. Role annotations for German have a different lineage: they originate from the FrameNet-annotated SALSA corpus (Burchardt et al., 2006) semi-automatically converted to PropBank style for the CoNLL-2009 shared task (Hajič et al., 2009), and enriched with VerbNet labels in SR3de (Mújdricza-Maydt et al., 2016). As a result, while English PropBank labels are assigned in a *predicate-independent* manner, German PropBank, following the same numbered labeling scheme,

---

[5] Darker color corresponds to higher similarity.

[6] Apart from the general tendency to use deeper layers in German reported in 5.1

property	R^2
(A) *instigation	0.68 (0.21)
(A) *volition	0.75 (0.11)
(A) *awareness	0.78 (0.09)
(A) *sentient	0.83 (0.07)
(A) *change.of.location	0.49 (0.04)
(A) *exists.as.physical	0.63 (0.03)
(P) *created	0.22 (0.01)
(P) *destroyed	0.11 (0.00)
(P) *changes.possession	0.26 (-0.01)
(P) *change.of.state	0.37 (0.01)
(P) *stationary	0.39 (0.05)

Table 6: Best dev R^2 for proto-role probing tasks over 20 epochs; A - Proto-Agent, P - Proto-Patient; Baseline in parentheses.

keeps this scheme *consistent within the frame*. We assume that this incentivizes the probe to learn semantic verb groupings and reflects in our probing results. The ability of the probe to detect subtle differences between formalism implementations constitutes a new use case for probing, and a promising direction for future studies.

5.5 Encoding of Proto-Roles

We now turn to the probing results for decompositional semantic proto-role labeling tasks. Unlike (Tenney et al., 2019b) who used a multi-label classification probe, we treat SPR properties as separate regression tasks. The results in Table 6 show that the performance varies by property, with some of the properties attaining reasonably high R^2 scores despite the simplicity of the probe architecture and the small dataset size. We observe that properties associated with Proto-Agent tend to perform better. The untrained mBERT baseline performs poorly which we attribute to the lack of data and the fine-grained semantic nature of the task.

Our fine-grained, property-level task design allows for more detailed insights into the layer utilization by the SPR probes (Fig. 4). The results indicate that while the layer utilization on the predicate side (`src`) shows no clear preference for particular layers (similar to the results obtained by Tenney et al. (2019a)), some of the proto-role features follow the pattern seen in the categorical role labeling and dependency parsing tasks for the argument tokens `tgt`. With few exceptions, we observe that the properties displaying that behavior
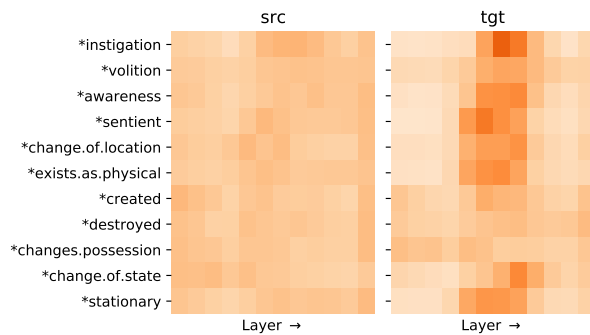
Figure 4: Layer utilization for SPR properties.

are Proto-Agent properties; moreover, a close examination of the results on syntactic preference by Reisinger et al. (2015, p. 483) reveals that these properties are also the ones with strong preference for the subject position, including the outlier case of `stationary` which in their data behaves like a Proto-Agent property. The correspondence is not strict, and we leave an in-depth investigation of the reasons behind these discrepancies for follow-up work.

6 Conclusion

We have demonstrated that the choice of linguistic formalism can have substantial, linguistically meaningful effects on role-semantic probing results. We have shown how probing classifiers can be used to detect discrepancies between formalism implementations, and presented evidence of semantic proto-role encoding in the pre-trained mBERT model. Our refined implementation of the edge probing framework coupled with the anchor task methodology enabled new insights into the processing of predicate-semantic information within mBERT. Our findings suggest that linguistic formalism is an important factor to be accounted for in probing studies.

This prompts several recommendations for the follow-up probing studies. First, the formalism and implementation used to prepare the linguistic material underlying a probing study should be always explicitly specified. Second, if possible, results on multiple formalisations of the same task should be reported and validated for several languages. Finally, assembling corpora with parallel cross-formalism annotations would facilitate further research on the effect of formalism in probing.

While our work illustrates the impact of formalism using a single task and a single probing framework, the influence of linguistic formalism per se is likely to be present for any probing setup that builds upon linguistic material. An investigation of how, whether, and why formalisms and their implementations affect probing results for tasks beyond role labeling and for frameworks beyond edge probing constitutes an exciting avenue for future research.

Acknowledgments

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.

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

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado. Association for Computational Linguistics.

Claire Bonial, Kevin Stowe, and Martha Palmer. 2013. Renewing and revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9–17. Association for Computational Linguistics.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:2126–2136.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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 Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 76(3):474–496.

Charles J. Fillmore. 1968. The Case for Case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 929–936, Sydney, Australia. Association for Computational Linguistics.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Karen Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. 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 2779–2795, Hong Kong, China. Association for Computational Linguistics.

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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv:1909.11942*.

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

Beth Levin and Malka Rappaport Hovav. 2005. *Argument Realization*. Research Surveys in Linguistics. Cambridge University Press.

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.

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

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK. Coling 2008 Organizing Committee.

Éva Mújdricza-Maydt, Silvana Hartmann, Iryna Gurevych, and Anette Frank. 2016. Combining semantic annotation of word sense & semantic roles: A novel annotation scheme for VerbNet roles on German language data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

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

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8594–8603. Curran Associates, Inc.

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. *arXiv:2002.12327*.

Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English universal dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).

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. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–17.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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

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, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv:1910.03771*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey,

Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*, abs/1609.08144.