Morph Call: Probing Morphosyntactic Content of Multilingual Transformers

Vladislav Mikhailov[1,2], **Oleg Serikov**[2,3], **Ekaterina Artemova**[2,4]
[1] SberDevices, Sberbank, Moscow, Russia
[2] HSE University, Moscow, Russia
[3] Neural Networks and Deep Learning Lab
Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[4] Huawei Noah's Ark lab, Moscow, Russia
Mikhaylov.V.Nikola@sberbank.ru {oserikov,elartemova}@hse.ru

Abstract

The outstanding performance of transformer-based language models on a great variety of NLP and NLU tasks has stimulated interest in exploring their inner workings. Recent research has focused primarily on higher-level and complex linguistic phenomena such as syntax, semantics, world knowledge, and common sense. The majority of the studies are anglocentric, and little remains known regarding other languages, precisely their morphosyntactic properties. To this end, our work presents **Morph Call**, a suite of 46 probing tasks for four Indo-European languages of different morphology: English, French, German and Russian. We propose a new type of probing task based on the detection of guided sentence perturbations. We use a combination of neuron-, layer- and representation-level introspection techniques to analyze the morphosyntactic content of four multilingual transformers, including their less explored distilled versions. Besides, we examine how fine-tuning for POS-tagging affects the model knowledge. The results show that fine-tuning can improve and decrease the probing performance and change how morphosyntactic knowledge is distributed across the model. The code and data are publicly available, and we hope to fill the gaps in the less studied aspect of transformers.

1 Introduction

In the last few years, transformer language models (Vaswani et al., 2017) have accelerated the growth in the field of NLP. The models have established new state-of-the-art results in multiple languages and even demonstrated superiority in NLU benchmarks compared to human solvers (Raffel et al., 2020; Xue et al., 2020; He et al., 2020). Their distilled versions, or so-called student models, have shown competitive performance on many NLP tasks while having fewer parameters (Tsai et al., 2019). However, many questions remain on how these models work and what they know about language. The previous research focuses on what knowledge has been learned during and after pre-training phases (Chiang et al., 2020; Rogers et al., 2020a), and how it is affected by fine-tuning (Gauthier and Levy, 2019; Peters et al., 2019; Miaschi et al., 2020; Merchant et al., 2020). Besides, a wide variety of language phenomena has been investigated including syntax (Hewitt and Manning, 2019a; Liu et al., 2019a), world knowledge (Petroni et al., 2019; Jiang et al., 2020), reasoning (van Aken et al., 2019), common sense understanding (Zhou et al., 2020; Klein and Nabi, 2019), and semantics (Ettinger, 2020).

Most of these studies involve *probing* which measures how well linguistic knowledge can be inferred from the intermediate representations of the model. The methods range from individual neuron analysis (Dalvi et al., 2020; Durrani et al., 2020a), examination of attention mechanisms (Kovaleva et al., 2019; Vig and Belinkov, 2019), correlation-based similarity measures (Wu et al., 2020), to probing tasks accompanied by linguistic supervision (Adi et al., 2016; Conneau et al., 2018).

Despite growing interest in interpreting the models, morphology has remained understudied, specifically for languages other than English. The majority of prior works on this subject are devoted to the introspection of machine translation models, word-level embedding models, or transformers, fine-tuned for POS-tagging (see Section 2).

To this end, we introduce **Morph Call**, a probing suite for the exploration of morphosyntactic content in transformer language models. The contributions of this paper are summarized as follows. First, we propose 46 probing tasks in four Indo-European languages of different morphology: Russian, French, English, and German. Inspired by techniques for model acceptability judgments

Proceedings of the Third Workshop on Computational Typology and Multilingual NLP, pages 97–121
June 10, 2021.

(Warstadt et al., 2019a) and adversarial training (Alzantot et al., 2018; Tan et al., 2020b,c), we present a new type of probing tasks based on the detection of guided sentence perturbations. Since the latter is automatically generated, the tasks can be adapted to other languages. Second, we use complementary probing methods to analyze four multilingual transformer encoders, including their distilled versions. We examine how fine-tuning for POS-tagging affects the probing performance and establish count-based and non-contextualized baselines for the tasks. Finally, we publicly release the tasks and code[1], hoping to fill the gaps in the less studied aspect of transformers.

2 Related Work

A large body of recent research is devoted to analyzing and interpreting the linguistic capacities of pre-trained contextualized encoders. The most common approach is to train a simple classifier for solving a probing task over the word- or sentence-level features produced by the models (Conneau et al., 2018; Liu et al., 2019a). The classifier's performance is used as a proxy to assess the model knowledge about a particular linguistic property. However, lately, the method has been critiqued: is the property truly learned by the model, or does the model encode the property for the classifier to easily extract it given the supervision? Besides, a new set of additional classifier parameters can make it challenging to interpret the results (Hewitt and Liang, 2019; Hewitt and Manning, 2019b; Saphra and Lopez, 2019; Voita and Titov, 2020).

Nevertheless, the probing classifiers are widely applied in the field of model interpretation, including morphology. One of the first works on morphological content is carried out on machine translation models where the classifier is learned to predict POS-tags in multiple languages (Belinkov et al., 2017, 2018). The latest studies involving POS properties in transformers show that they are predominantly captured at the lower layers (Tenney et al., 2019b; Liu et al., 2019b; Rogers et al., 2020a), and can be evenly distributed across all layers (Durrani et al., 2020b). *Amnesic* probing explores how removing information at a particular layer affects the probe performance at the final layer (Elazar et al., 2020). This allows measuring the layer importance with respect to a linguistic property. The results claim that removing POS information may affect the performance more at the higher layers as compared to the lower ones.

Another line of research is devoted to various linguistic phenomena at the juxtaposition of morphology, syntax, and semantics. LSTM-based models and transformers are probed to capture subject-verb agreement in different languages (Linzen et al., 2016; Giulianelli et al., 2018; Ravfogel et al., 2018; Goldberg, 2019). Recently, the agreement has been at the core of inflectional perturbations for adversarial training (Tan et al., 2020a), and linguistic acceptability judgments along with morphological, syntactic, and semantic violations (Warstadt et al., 2019b).

Our work is closely related to (Edmiston, 2020) who explore morphological properties and subject-verb agreement in the hidden representations and self-attention heads of transformer models. However, there are several differences. First, we investigate the knowledge in multilingual transformers and their distilled versions instead of monolingual ones. Second, we carry out the experiments on an extended set of tasks, such as detecting syntactic and inflectional perturbations (see Section 3.2). Third, we apply several probing methods to analyze from different perspectives. Finally, we study the impact of fine-tuning for POS-tagging on the probe performance. Despite the similarities and differences, we find the studies complementary.

Finally, such benchmarks as LINSPECTOR (Şahin et al., 2020) and XTREME (Hu et al., 2020) provide means for evaluation of multilingual embedding models and cross-lingual transferring methods with regards to multiple linguistic properties, specifically morphology.

3 Method

3.1 Morphosyntactic Inventories

This paper investigates four Indo-European languages that fall under different morphological types: Russian, French, English, and German. Russian and French have fusional morphology, while English is an analytic language, and German exhibits peculiarities of fusional and agglutinative types. We consider the nominal morphosyntactic features of Number, Case, Person, and Gender. Even though the feature inventory is mostly shared across the languages, the latter differ significantly in their richness of morphology (Baerman, 2007).

[1] `https://github.com/morphology-probing/morph-call`

The morphosyntactic inventories of the analyzed languages are outlined in Table 1.

3.2 Probing Tasks

Data We use sentences from the Universal Dependencies (UD) (Nivre et al., 2016) for all our probing tasks, keeping in mind possible inconsistency between the Treebanks (de Marneffe et al., 2017; Alzetta et al., 2017; Droganova et al., 2018), and consequent inconsistency in dataset sizes across languages. All sentences are filtered by a 5-to-25 token range, and each task is split into 80/10/10 train/val/test partitions with no sentence overlap. The partitions are balanced by the number of instances per target class. Notably, the availability of the UD Treebanks in different languages allows for an adaptation of the method to the other ones. The used Treebanks are listed in Appendix A, and a brief statistics of the tasks is presented in Appendix B.

Task Description We construct four groups of probing tasks framed as binary or multi-class classification tasks: **Morphosyntactic Features**, **Masked Token**, **Morphosyntactic Values** and **Perturbations**.

Morphosyntactic Features probe the encoder for the occurrence of the morphosyntactic properties. The goal is to detect if a word exhibits a particular property based on its contextualized representation. Consider an example for the Russian sentence *'The clock stopped in a month.'*:

$$\text{Chasy } \underbrace{\textbf{ostanovilis'}}_{\text{to stop+3PL+PST (1)}} \; \underbrace{\textbf{cherez}}_{\text{in (0)}} \text{ mesyats .}$$

Here, the target words are indicated by bold, and the labels denote if they have the category of Number.

Masked Token tasks are analogous to **Morphosyntactic Features** with the exception that the target word is replaced with a tokenizer-specific mask token. The tasks test if it is possible to recover the properties of the masked token purely from the context. Below is an example where the sentence mentioned above *'The clock stopped in a month.'* contains masked target words, and labels denote the occurrence of the Number feature at the position of the token:

$$\text{Chasy } \underset{1}{\textbf{[MASK]}} \text{ cherez mesyats .}$$

$$\text{Chasy ostanovilis' } \underset{0}{\textbf{[MASK]}} \text{ mesyats .}$$

Morphosyntactic Values is a group of k-way classification tasks for each feature where k is the number of values that the feature can take (see Table 1). For instance, the goal is to identify whether the word *girl* is in the singular or plural form: 'The **girl** has either pink or brown.'

Perturbations tasks test the encoder sensitivity to various sentence perturbations. Removing words from a text has recently been used to obtain adversarial attacks (Liang et al., 2017; Li et al., 2018), whereas inflectional perturbations have been applied for adversarial training of transformers (Tan et al., 2020b,c). In contrast, we extend the perturbations to probe the encoders for linguistic knowledge. To this end, we construct eight tasks that involve syntactic perturbations and inflectional perturbations in the subject-predicate agreement and deictic words. Note that we apply a set of language-specific rules to control the quality of the error generation procedure. To obtain the inflectional candidates, we make use of pymorphy2 for Russian (Korobov, 2015), lemminflect[2] for English, and word paradigm tables from Wiktionary for French[3] and German[4].

Stop-words Removal involves corruption of a syntax tree by removing stop-words. We use lists of stop-words provided by NLTK library (Loper and Bird, 2002). Consider an example of the French sentence *'**Les** Irakiens **ont** tout détruit **à le** Koweit'*, where the bolded words correspond to the removed stop-words.

Article Removal is a special case of the previous task, revealing whether the encoders are sensitive to discarded articles. This task is only constructed for French, English, and German. Note that such perturbation may also strain the semantics of the sentence: *'It's on loan, by **the** way'*.

Subject Number includes inflectional perturbations of the subject in the main clause with respect to the Number: *'The **girls** has either pink or brown.'*

[2] `https://github.com/bjascob/LemmInflect`

[3] `https://dumps.wikimedia.org/frwiktionary/latest/`

[4] `https://dumps.wikimedia.org/dewiktionary/latest/`

Feature \ Language	English	French	German	Russian
Number	$\{Sing, Plur\}$	$\{Sing, Plur\}$	$\{Sing, Plur\}$	$\{Sing, Plur\}$
Case	–	–	$\{Nom, Acc, Dat, Gen\}$	$\{Nom, Acc, Dat, Gen, Loc, Ins\}$
Person	$\{1, 2, 3\}$	$\{1, 2, 3\}$	$\{1, 2, 3\}$	$\{1, 2, 3\}$
Gender	–	$\{Masc, Fem\}$	$\{Masc, Fem, Neut\}$	$\{Masc, Fem, Neut\}$

Table 1: Analyzed languages and their morphosyntactic feature inventories.

Subject Case comprises errors in the subject form of Case for Russian. Consider an example of the perturbed sentence *Kak **vy** vidite situatsiyu v Rossii?* 'How do you find the situation in Russia?', where the nominative form of the subject *vy* 'you' is changed to the accusative:

$$\text{Kak } \underbrace{\textbf{vas}}_{\text{you+2PL+ACC}} \text{ vidite situatsiyu v Rossii ?}$$

Predicate Number incorporates perturbations of the predicate in the main clause regarding the Number feature: *'It **make** a huge difference.'*

Predicate Gender contains errors in the Gender form of the predicate in the main clause. For example, the masculine form of the predicate *byl* 'was' in the Russian sentence *Dosug **byl** ves'ma odnoobrazen* 'The leisure was pretty monotonous' is changed to the feminine:

$$\text{Dosug } \underbrace{\textbf{byla}}_{\text{to be+3SG+FEM}} \text{ ves'ma odnoobrazen .}$$

Predicate Person comprises perturbations in the Person form of the predicate in the main clause. For instance, the Russian sentence *Ya **poedu** v Moskvu* 'I will go to Moscow' contains the perturbed predicate in the form of the second Person instead of the first one:

$$\text{Ya } \underbrace{\textbf{poedesh'}}_{\text{to go+2SG}} \text{ v Mosckvu .}$$

Deictic Word Number involves perturbations generated by the inflection of demonstrative pronouns (only in English and German). For example, the singular form of the pronoun *dieser* 'this' is changed to the plural form *diesen* 'these' in the sentence *Siehe zu **dieser** Technik auch* 'See also this technique':

$$\text{Siehe zu } \underbrace{\textbf{diesen}}_{\text{this+PL+DAT}} \text{ Technik auch .}$$

4 Experimental Setup

4.1 Models

The experiments are run on the following multilingual transformer models released as a part of HuggingFace library (Wolf et al., 2019):

M-BERT (Devlin et al., 2019) was pre-trained over concatenated monolingual Wikipedia corpora in 104 languages.

D-BERT (Sanh et al., 2019) or DistilBERT is a 6-layer distilled version of **M-BERT** model.

XLM-R (Conneau et al., 2019) was pre-trained over filtered CommonCrawl data in 100 languages (Wenzek et al., 2019).

MiniLM (Wang et al., 2020) is a distilled **M-BERT** model that uses **XLM-R** tokenizer.

Each model under investigation has two instances for each language:

1. *Fine-tuned model* is a transformer model fine-tuned for POS-tagging. We use the UD Treebanks and HuggingFace library for fine-tuning. The data is randomly split into 80/10/10 train/val/test sets.

2. *Pre-trained model* is a non-tuned transformer model with frozen weights.

4.2 Probing Methods

Probing Classifiers We use Logistic Regression from scikit-learn library (Pedregosa et al., 2011) as a probing classifier. The classifier is trained over hidden representations[5] produced by the encoders with the regularization parameter $L^2 \in [0.25, 0.5, 1, 2, 4]$ tuned on the validation set. The performance is evaluated by the ROC-AUC score.

[5]Morphosyntactic Features and Values: we take mean-pooled representations of the sub-word embeddings that correspond to a target word. Masked Token: we use embedding of a tokenizer-specific masked token. Perturbations: we use mean-pooled sentence representations.

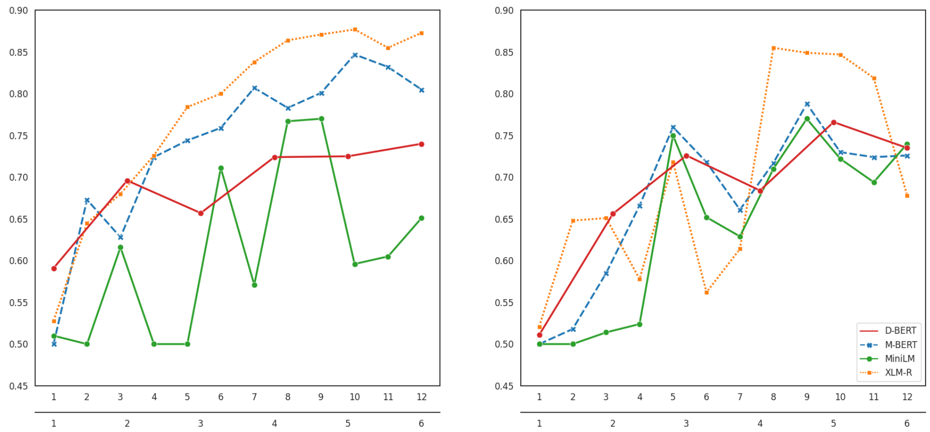

Figure 1: The performance of the probing classifier on **Case** masked token task for Russian. X-axis=Layer index score. Y-axis=Accuracy score. Left: pre-trained models. Right: fine-tuned models.

Neuron Analysis The neuron-level analysis allows retrieving a group of individual neurons that are most relevant to predict a linguistic property (Durrani et al., 2020a). Similarly, a linear classifier is trained over concatenated mean-pooled word/sentence representations using Elastic-net regularization (Zou and Hastie, 2005), and with L^1 and L^2 λ's $\in [0.1, \ldots, 1e^{-5}]$ tuned on the validation set. The weights of the classifier are used to measure the relevance of each neuron.

Correlation Analysis Canonical correlation analysis (`ckasim`) is a representation-level similarity measure that allows identifying pairs of layers of similar behavior (Wu et al., 2020). We use `[CLS]`-pooled intermediate representations to analyze the encoders. The measure is computed with the help of the publicly available code[6].

4.3 Baselines

We train Logistic Regression over the following count-based and distributive baseline features (see Section 4.2). We use N-gram range $\in [1, 4]$ for each count-based baseline. Count-based features include **Char Number** (length of a word/sentence in characters), **TF-IDF over character N-grams**, **TF-IDF over BPE tokens** (BertTokenizer), and **TF-IDF over SentencePiece tokens** (XLMRobertaTokenizer). We use multilingual tokenizers by HuggingFace library to split words/sentences into the sub-word tokens. The distributive baseline is mean-pooled monolingual **fastText**[7] word/sentence embeddings (Bojanowski et al., 2017).

5 Results

5.1 Morphosyntactic Features

Probing Classifiers We learn the probing classifiers to estimate the model awareness of the morphosyntactic properties (see Section 4.2). The results demonstrate that pre-trained models perform slightly worse than their fine-tuned versions (2-4%). We find that the awareness is distributed in a very similar manner despite the language differences, for the models of both instances (see Tables 6–7, Appendix D). Specifically, the performance on **Number** and **Gender** is reaching its plateau at the middle layers $[5-8]$ of 12-layer models, and at layer $[3]$ of **D-BERT**. The probing curves[8] on **Case** are achieving their peak at the lower-to-middle layers $[4-5]$ and staying at the plateau towards the output layer. The only difference is observed on **Person** where the property is best inferred either across all layers (English, Russian) or at the lower-to-higher layers $[4-11]$ (French, German). The baseline features receive a strong performance, meaning that the occurrence of certain property may be inferred using the sub-word information (see Table 4, Appendix C).

[6] `https://github.com/johnmwu/contextual-corr-analysis`

[7] `https://fasttext.cc/docs/en/crawl-vectors.html`

[8] We refer to a *probing curve* as to a graphical representation of the probing classifier performance.

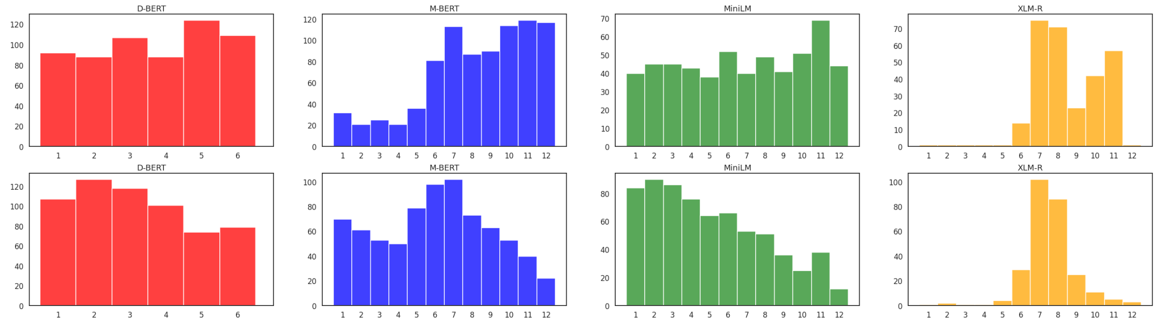

Figure 2: The distribution of top neurons over **Predicate Gender** perturbation task for each model. X-axis=Layer index number. Y-axis=Number of neurons. Top: pre-trained models. Bottom: fine-tuned models.

5.2 Masked Token

Probing Classifiers The results of the probing classifier performance on **Masked Token** tasks are presented in Tables 8 (pre-trained models) and 9 (fine-tuned models) (see Appendix D). The task has appeared to be more challenging as opposed to **Morphosyntactic Features** (see Section 5.1). An interesting observation in this setting is that the performance of the models predominantly drops or becomes unstable after fine-tuning. For instance, **BERT** may lose almost 10% in the tasks for Russian, and **D-BERT** may drop 5% in the tasks for French. The probing curves tend to show rapid increases and decreases across the layers. An exception to this pattern is **XLM-R** which is less affected by fine-tuning and exhibits a more stable probing behavior. Nevertheless, the models demonstrate their capability to infer the properties from the context. **XLM-R** makes correct predictions in almost 70% of cases, while the performance of **M-BERT** and **D-BERT** is slightly worse, and **MiniLM** may struggle the most. Figure 1 outlines the results on **Case** task for Russian, best solved among the others. The middle-to-higher layers account for more correct predictions in the models of both instances. However, the higher layers $[10-12]$ of 12-layer models and layer $[6]$ of **D-BERT** may pertain to lower performance. A possible explanation is that the layers are affected by the objectives, i.e., Masked Language Modeling (pre-trained) or POS-tagging (fine-tuned). We find that the contextualized representations of a masked token produced by the final layers of pre-trained models may store the morphosyntactic properties. The probing curves demonstrate that the distribution of the properties may get affected by fine-tuning, or the knowledge can be partially lost, which is shown by the performance drops.

5.3 Morphosyntactic Values

Property-wise Neuron Analysis We apply property-wise neuron analysis to investigate the top-neurons per each morphosyntactic property (see Section 4.2). We find that some models require a larger group of neurons to learn a morphosyntactic property, and the number of these neurons may get changed after fine-tuning. We provide the results for each language in Appendix E. Figure 4 illustrates the distributions for pre-trained and fine-tuned models for French. While after fine-tuning the number of neurons on **Person** (**M-BERT**, **D-BERT**) and **Number** (**XLM-R**) has increased, **Number** and **Gender** are now handled by fewer neurons of the distilled models (**D-BERT**, **MiniLM**). A similar behavior is observed for Russian and English. **Case** (Russian), **Gender** (Russian) and **Person** (English) require more neurons (**M-BERT**), or fewer neurons over **Person** (Russian) and **Number** (Russian, English) (**MiniLM**, **D-BERT**). Notably, the fine-tuning phase does not affect the neuron distributions for German.

5.4 Perturbations

Probing Classifiers The results of the probing classifier performance on **Perturbations** tasks are presented in Table 10 (pre-trained models), and Table 11 (fine-tuned models) (see Appendix D). We find that the models perform on par with one another in the majority of the tasks. Notably, **XLM-R** is generally the most sensitive to the perturbations in each language compared to the other models. We find that the syntactic perturbations (**Article Removal**, **Stopwords Removal**) are better solved than the inflectional ones. Similarly, the count-based baselines receive the best performance on the syntactic perturbations since the latter are obtained over a limited set of words (see Table 5,

Appendix C). On the other hand, their performance is typically higher or close to random on the inflectional perturbations (see Table 5, Appendix C). We briefly describe the results in Appendix D for the sake of space.

Layer-wise Neuron Analysis Individual neuron analysis helps to observe how top-neurons are spread across the entire model, and identify the relevance of each layer by the number of its top-neurons[9] (see Section 4.2). Figure 2 demonstrates the results for **Predicate Gender** task in Russian. The sensitivity to the perturbation tends to be distributed across all layers of both pre-trained and fine-tuned models (**D-BERT**, **M-BERT**, **MiniLM**). The exception is provided by **XLM-R** which localizes the knowledge at the middle-to-higher layers $[6-11]$ (pre-trained), or in fewer layers but with larger groups of neurons $[6-9]$ (fine-tuned). The models of both instances store the sensitivity to the incorrect subject case form (**Subject Case**, Appendix E) at the middle-to-higher layers (**D-BERT**: $[3-5]$, **M-BERT**: $[6-11]$, **MiniLM**: $[4-8]$, **XLM-R**: $[5-12]$). Notably, the number of top-neurons in all models has decreased after the fine-tuning, and the information has been now more localized in two of them (**MiniLM**, **XLM-R**). A similar behavior of the models by language is observed on **Subject Number** (see Appendix E). The property is generally captured at the middle-to-higher layers of each pre-trained model for Russian, German and French (**D-BERT**: $[2, 3-6]$, **M-BERT**: $[6-12]$, **MiniLM**: $[5-12]$). The results are different for their fine-tuned versions, where the property gets more localized for Russian and German (**D-BERT**: $[3-5]$, **MiniLM**: $[5-7]$, **XLM-R**: $[6-9]$, **M-BERT**: $[6-11]$), or captured by fewer neurons at the same layers for French. In contrast, the property is predominantly distributed across all layers of both pre-trained and fine-tuned models for English.

Correlation Analysis To analyze the encoders with `ckasim`, we take `[CLS]`-pooled representations of the original sentence (without the perturbation) and its perturbed version. The similarity measure is computed on the resulted pairs of representations. For each model **M** we explore three settings by combining different model instances (see Section 4.1): (i) (*pre-trained* **M**, *pre-trained* **M**), (ii) (*pre-trained* **M**, *fine-tuned* **M**), (iii) (*fine-tuned* **M**, *fine-tuned* **M**). Figure 3 shows the most typical pattern achieved in the tasks. The biggest difference is observed over the combination (ii), where the perturbations are best captured at the lower-to-middle layers $[1-6]$ (**XLM-R**, **MiniLM**), or across all the layers (**M-BERT**, **DistilBERT**). The middle-to-higher layers $[7-12]$ tend to become more similar over combinations (i, iii) which may mean that they are able to restore the semantics of the perturbed sentences, being more robust to the perturbations as opposed to the lower ones.

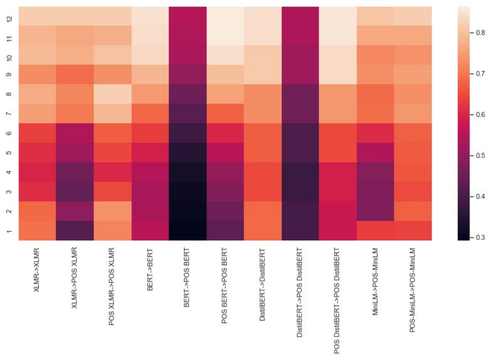

Figure 3: `ckasim` results on **Stopwords Removal** task in German. X-axis=Model instance combinations. Y-axis=Layer index number (left), `ckasim` score (right).

6 Discussion

Morphosyntactic content across languages The probing curves under layer-wise probing demonstrate that the multilingual transformers learn the morphosyntactic content in a greatly similar manner despite the language differences (see Section 5.1). The properties are predominantly distributed across the middle-to-higher layers $[5-12]$ for each language. In contrast, **Masked Token** tasks represent a challenge for the models causing rapid increases and decreases in the performance across the layers (see Section 5.2). The overall pattern for each language is that a masked token's properties are best inferred at the middle-to-higher layers. A possible reason for this is that the task requires incorporating syntactic and semantic information from the context since the target word remains unseen. The models demonstrate their sensitivity to **Perturbations** (see Section 5.4). While the syntactic perturbations are predominantly captured at the lower-to-middle layers $[3-8]$, the inflectional ones are stored at the middle-to-higher

[9]We selected top-20% neurons using the neuron ranking algorithm (Durrani et al., 2020b).

layers $[5 - 12]$. In contrast to other languages, the perturbation properties for English may be distributed across all layers of the models. The results are supported by the individual neuron analysis, an example of which is provided in Appendix E.

Same properties require different number of neurons Property-wise neuron analysis shows that **Person** and **Case** are learned using more neurons as compared to **Number** and **Gender** across the languages. Notably, the number of neurons required to learn a property may depend on the language. For example, **D-BERT** requires about 1000 neurons to learn **Case** in German and less than 1500 neurons to learn the property in Russian.

Are students good learners? A common method to compare pre-trained models and their distilled versions is based upon their performance on downstream tasks (Tsai et al., 2019), or NLU benchmarks (Wang et al., 2018, 2019). Still, little is investigated on what language properties are preserved after the knowledge distillation. We find that **D-BERT** and **MiniLM** mimic the behavior of their teachers under layer-wise probing (see Section 5.1), or display a similar perturbation sensitivity under `ckasim` (see Section 5.4). However, **MiniLM** tends to exhibit an uncertain behavior as opposed to their teacher (see Sections 5.2, 5.4).

Effect of fine-tuning The results show that the effect of fine-tuning for POS-tagging varies within a certain group of tasks. First, fine-tuned models may receive a better probing performance by 2-4% on **Morphosyntactic Features** tasks (see Section 5.1). Second, fine-tuning affects the way the properties are distributed or causes significant performance drops on **Masked Token** tasks, specifically at the higher layers (see Section 5.2). The impact on the property distribution is also demonstrated on **Perturbations** tasks under neuron-level probe (see Section 5.4). Besides, the analysis of top-neurons allows concluding that fine-tuning may affect localization (**MiniLM**, **XLM-R**) which is in line with (Wu et al., 2020). Finally, a number of neurons required to predict a property may increase (e.g., Russian: **Case**; French: **Person**), decrease (e.g., English: **Number**) or remain unchanged (German). We suggest that an interesting line for future work is to analyze the correlation between the number of neurons and the probe performance after fine-tuning. For instance, the results on **Perturbation** tasks indicate that some models may receive a better probing performance with fewer (**XLM-R**) or more neurons (**D-BERT**, **M-BERT**) (see Section 5.4). An exploration of fine-tuning for morphosyntactic analysis, specifically over UniMorph (Kirov et al., 2018) may be a fruitful avenue for future work.

Distribution of knowledge may depend on language morphology The analysis of the models under layer-wise and neuron-wise probing suggests that the behavior may depend on how morphologically rich a language is (see Sections 5.1, 5.4). The knowledge for English tends to be distributed across all layers of the models in contrast to the more morphologically rich languages that capture the properties at the middle-to-higher layers. The finding is in line with a few recent studies (Edmiston, 2020; Durrani et al., 2020b; Elazar et al., 2020) which contradict the common understanding that morphology is stored at the lower layers (Tenney et al., 2019a; Rogers et al., 2020b). We also find that the distribution of the properties varies based on the complexity of a probing task (see Sections 5.1, 5.2). An exciting direction for future work is to test this hypothesis on a more diverse set of morphologically contrasting languages. Besides, perturbing one aspect of a sentence can cause ambiguity elsewhere which is an interesting line for future exploration of the interdependence of the perturbations.

7 Conclusion

This paper proposes **Morph Call**, a suite of 46 probing tasks in four Indo-European languages that differ significantly in their richness of morphology: Russian, French, English, and German. The suite includes a new type of probing task based on the detection of syntactic and inflectional sentence perturbations. We apply a combination of three introspection methods based on neuron-, layer- and representation-level analysis to probe five multilingual transformer models, including their less explored distilled versions. The analysis of transformers' understudied aspect contradicts the common findings on how morphology is represented in the models. We find that the knowledge for English is predominantly distributed across all layers of the models in contrast to more morphologically rich languages (German, Russian, French), which house the properties at the middle-to-higher layers. The models demonstrate their sensitivity to the perturbations, and **XLM-R** tends to be the most robust among the others. We observe that distilled models inherit their teachers' knowledge, showing a comparative performance and exhibiting similar property distribution on several probing tasks. Another finding is that fine-tuning for POS-tagging can affect the model knowledge in various manners, ranging from improving and decreasing the probing classifier performance to changing the information's localization. We believe there is still room for exploring the models' morphosyntactic content and the effect of fine-tuning, specifically across a more diverse set of languages and types of model architectures.

Acknowledgements

We thank our reviewers for their insightful comments and suggestions. Ekaterina Artemova is supported by the framework of the HSE University Basic Research Program.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. *arXiv preprint arXiv:1608.04207*.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 201–210.

Matthew Baerman. 2007. Syncretism. *Language and Linguistics Compass*, 1(5):539–551.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? *arXiv preprint arXiv:1704.03471*.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv preprint arXiv:1801.07772*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained Language Model Embryology: The Birth of ALBERT. pages 6813–6828.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing Redundancy in Pretrained Transformer Models. pages 4908–4926.

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the universal dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115.

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.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*, pages 52–65.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020a. Analyzing Individual Neurons in Pre-trained Language Models. pages 4865–4880.

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

Daniel Edmiston. 2020. A systematic analysis of morphological content in bert models for multiple languages. *arXiv preprint arXiv:2004.03032*.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. When bert forgets how to pos: Amnesic probing of linguistic properties and mlm predictions. *arXiv preprint arXiv:2006.00995*.

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.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. 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 529–539.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*.

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

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

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.

John Hewitt and Christopher D. Manning. 2019a. 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.

John Hewitt and Christopher D Manning. 2019b. A Structural Probe for Finding Syntax in Word Representations. pages 4129–4138.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. UniMorph 2.0: Universal Morphology.

Tassilo Klein and Moin Nabi. 2019. Attention is (not) all you need for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4831–4836.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 320–332. Springer.

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.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2017. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*.

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

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

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

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

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. *the Journal of machine Learning research*, 12:2825–2830.

Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Shauli Ravfogel, Francis M Tyers, and Yoav Goldberg. 2018. Can lstm learn to capture agreement? the case of basque. *arXiv preprint arXiv:1809.04022*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020a. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

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

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics*, 46(2):335–385.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. 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 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020a. It's morphin' time! Combating linguistic discrimination with inflectional perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2920–2935, Online. Association for Computational Linguistics.

Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher. 2020b. It's morphin'time! combating linguistic discrimination with inflectional perturbations. *arXiv preprint arXiv:2005.04364*.

Samson Tan, Shafiq Joty, Lav Varshney, and Min-Yen Kan. 2020c. Mind Your Inflections! Improving NLP for Non-Standard Englishes with Base-Inflection Encoding. pages 5647–5663.

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, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and Practical BERT Models for Sequence Labeling. pages 3623–3627.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1823–1832.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. pages 5998–6008.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. *arXiv preprint arXiv:2003.12298*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-purpose Language Understanding Systems. pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

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

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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

John M Wu, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2020. Similarity analysis of contextual word representation models. *arXiv preprint arXiv:2005.01172*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *AAAI*, pages 9733–9740.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Appendix

A Description of Treebanks

Below is a list of the UD Treebanks used in the experiments:

- **Russian**: GramEval2020 Treebanks, GSD Russian Treebank, Russian-PUD, and SynTagRus Treebank.
- **English**: EWT Treebank, GUM Treebank, the English portion of ParTUT, English-PUD, and English-Pronouns Treebank.
- **French**: French Question Bank, GSD French Treebank, the French portion of ParTUT, French-PUD, Sequoia and French Spoken Treebank, adapted from the Rhapsoide prosodic-syntactic Treebank.
- **German**: GSD German Treebank, HDT-UD Treebank, German-PUD and LIT German Treebank.

B Dataset Statistics

Tables 1 – 3 provide a brief statistics on the partition sizes for each probing task.

Probing Task	Language	Train	Dev	Test	Overall
Number	**Ru**	174 720	21 937	21 379	218 036
	En	51 465	6492	6374	64 331
	De	533 898	66 984	66 984	668 271
	Fr	74 450	9385	9191	93 026
Case	**Ru**	174 884	21 768	21 974	218 626
	De	436 303	54 692	53 932	544 927
Person	**Ru**	162 345	20 313	20 319	202 977
	En	47 001	5945	5735	58 681
	De	471 132	58 847	58 438	588 417
	Fr	71 394	8853	8992	89 239
Gender	**Ru**	165 934	20 462	20 982	207 378
	De	500 628	62 163	62 612	625 403
	Fr	69 901	8840	8559	87 300

Table 1: Number of samples for each **Morphosyntactic Features** and **Masked Token** task. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Probing Task	Language	Train	Dev	Test	Overall
Number	**Ru**	100 738	12 592	12 593	125 923
	En	21 568	2696	2696	26 960
	De	339 744	42 468	42 468	424 680
	Fr	33 339	4167	4168	41 674
Case	**Ru**	92 320	11 540	11 540	115 400
	De	252 182	31 523	31 523	315 228
Person	**Ru**	15 748	11 540	11 540	19 685
	En	7255	907	907	9069
	De	184 788	23 099	23 099	230 986
	Fr	6364	796	796	7956
Gender	**Ru**	76 158	9520	9520	95 198
	De	252 182	31 523	31 523	315 228
	Fr	23 660	2957	2958	29 575

Table 2: Number of samples for each **Morphosyntactic Values** task. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Probing Task	Language	Train	Dev	Test	Overall
Stop-words Removal	**Ru**	38 838	4855	4855	48 548
	En	12 627	1578	1578	15 784
	De	121 272	15 159	15 159	151 590
	Fr	13 959	1745	1745	17 449
Article Removal	**En**	7770	971	972	15 784
	De	99 669	12459	12459	124 587
	Fr	10 083	1253	1276	12 612
Subject Number	**Ru**	9293	1164	1165	11 622
	En	471	58	60	589
	De	5 709	1007	1009	6005
	Fr	1219	151	153	1523
Subject Case	**Ru**	18 897	2344	2346	23 587
Predicate Number	**Ru**	7160	897	897	8 954
	En	1115	140	142	1397
	De	26 415	4374	4375	35 164
	Fr	2822	353	356	3531
Predicate Person	**Ru**	5240	644	646	6530
Predicate Gender	**Ru**	4414	550	553	5517
Deixis Word Number	**En**	1130	141	142	1413
	De	4804	600	601	6005

Table 3: Number of samples for each **Perturbation** task. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

C Baseline Performance

Table 4 summarizes the results of the baseline models for **Morphosyntactic Features** tasks. Table 5 presents the performance of the baseline models for **Perturbations** tasks.

Probing Task	Lang	Char Num	TF-IDF Char	TF-IDF BPE	TF-IDF SP	fT
Number	**Ru**	0.78	**0.97**	0.96	0.96	0.94
	En	0.63	**0.95**	0.94	**0.95**	0.93
	De	0.57	**0.95**	**0.95**	0.95	0.89
	Fr	0.52	**0.91**	**0.91**	**0.91**	0.87
Case	**Ru**	0.69	**0.97**	0.96	0.96	0.90
	De	0.64	0.92	**0.93**	0.92	0.88
Person	**Ru**	0.60	**0.98**	**0.98**	**0.98**	0.93
	En	0.62	0.97	0.97	0.97	**0.98**
	De	0.66	**0.93**	**0.93**	**0.93**	0.91
	Fr	0.54	**0.93**	0.92	0.92	0.88
Gender	**Ru**	0.73	**0.96**	0.95	**0.96**	0.89
	De	0.47	**0.86**	**0.86**	**0.86**	0.81
	Fr	0.54	**0.88**	**0.88**	0.87	0.84

Table 4: Baseline results on **Morphosyntactic Features** tasks. **SP** refers to SentencePiece, and **fT** corresponds to fastText. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Probing Task	Lang	Char Num	TF-IDF Char	TF-IDF BPE	TF-IDF SP	fT
Stop-words Removal	**Ru**	0.57	**0.96**	0.92	0.92	0.93
	En	0.64	0.97	**0.98**	0.97	0.96
	De	0.63	**0.99**	**0.99**	**0.99**	0.97
	Fr	0.60	**0.98**	**0.98**	**0.98**	0.96
Article Removal	**En**	0.52	0.98	**0.99**	0.98	0.84
	De	0.55	**0.97**	**0.97**	**0.97**	0.87
	Fr	0.56	0.95	**0.97**	0.96	0.87
Subject Number	**Ru**	0.50	0.54	**0.55**	0.54	0.53
	En	**0.43**	0.35	0.37	**0.43**	0.40
	De	0.5	0.48	0.46	0.48	**0.57**
	Fr	0.44	**0.60**	0.50	0.55	0.55
Subject Case	**Ru**	0.51	**0.67**	0.62	0.62	0.60
Predicate Number	**Ru**	0.49	**0.64**	0.48	0.50	0.52
	En	**0.52**	0.49	0.45	0.47	0.48
	De	0.50	0.60	0.39	0.38	**0.68**
	Fr	0.49	0.64	0.47	0.49	**0.68**
Predicate Person	**Ru**	0.50	**0.81**	0.78	0.74	0.62
Predicate Gender	**Ru**	0.50	**0.62**	0.57	0.58	0.51
Deixis Word Number	**En**	0.48	0.71	**0.77**	0.75	0.70
	De	0.49	0.68	0.71	**0.72**	0.62

Table 5: Baseline results on **Perturbation** tasks. **SP** refers to SentencePiece, and **fT** corresponds to fastText. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

D Probing Classifiers

Morphosyntactic Features Tables 6 – 7 summarize the results of the probing classifier on **Morphosyntactic Features** tasks for pre-trained and fine-tuned models. Figure 1 shows a few examples of the model behavior on the tasks. While **Gender** in German appears to be the most challenging property among the others for both pre-trained and fine-tuned models, **Case** in Russian is inferred by the models with great confidence.

Masked Token Tables 8 – 9 outline the performance of the probing classifier on **Masked Token** tasks.

Perturbations Tables 10 – 11 present the results of the probing classifier on **Perturbations** tasks for pre-trained and fine-tuned models. Figures 2 – 3 are the graphical representations of the probing classifier performance on **Article Removal** task for German, and **Predicate Number** task for French.

The overall pattern for the syntactic perturbations is that the sensitivity is captured at the lower-to-middle layers $[3 - 8]$ of pre-trained models. In its turn, the inflectional properties are predominantly distributed at the middle-to-higher layers $[5 - 12]$ of both pre-trained and fine-tuned models. However, fine-tuned versions may exhibit unpredictable behavior, an example of which we describe below. Figure 2 demonstrates the results on **Article Removal** task for German. While the probing curves of pre-trained models tend to be decaying after reaching their peak at the middle layers, they are confidently increasing towards the output layer after the fine-tuning phase. In contrast, a different behavior is observed on **Predicate Number** task for French (see Figure 3). The layers of many fine-tuned models lose their knowledge (**MiniLM**: $[5 - 12]$, **D-BERT**: $[5]$, **M-BERT**: $[6 - 11]$, **XLM-R**: $[7; 11 - 12]$).

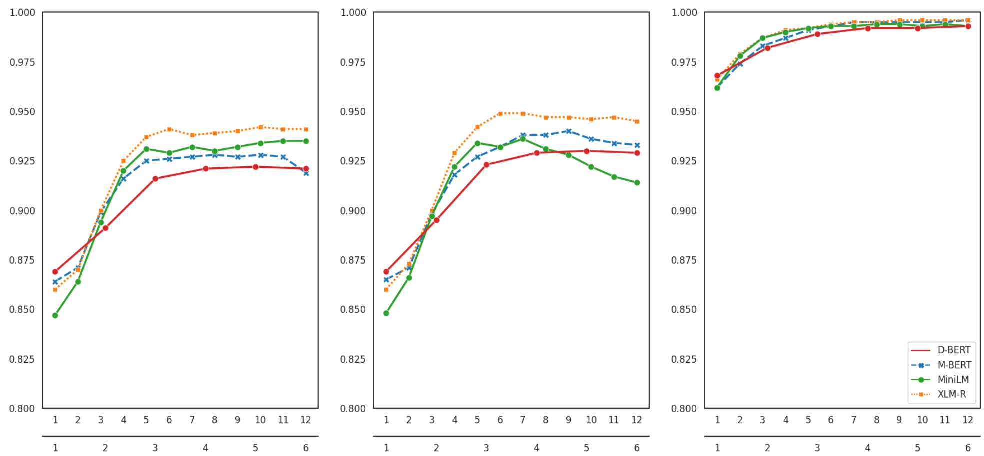

Figure 1: The performance of the probing classifier on **Morphosyntactic Features** tasks. Left: **Gender** in German (pre-trained). Middle: **Gender** in German (fine-tuned). Right: **Case** in Russian (fine-tuned).

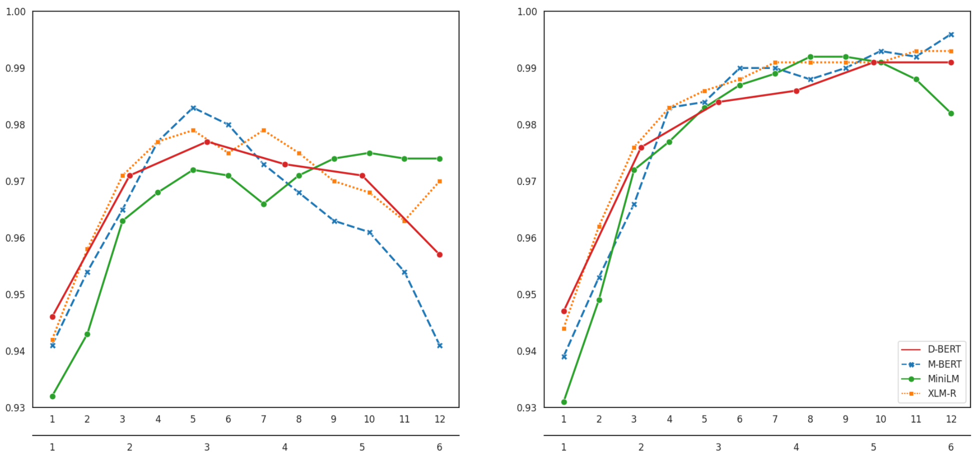

Figure 2: The performance of the probing classifier on **Article Removal** perturbation task for German. X-axis=Layer index number. Y-axis=Accuracy score. Left: pre-trained models. Right: fine-tuned models.

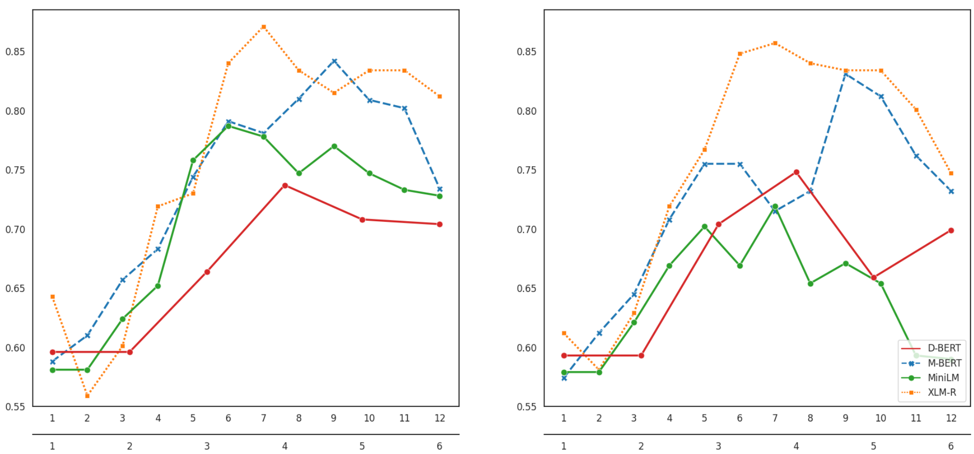

Figure 3: The performance of the probing classifier on **Predicate Number** perturbation task for French. X-axis=Layer index number. Y-axis=Accuracy score. Left: pre-trained models. Right: fine-tuned models.

Lang	Probing Task	D-BERT	MiniLM	BERT	XLM-R
De	**Case**	0.89	**0.91**	0.89	—
	Gender	0.91	**0.92**	0.91	**0.92**
	Number	0.93	**0.94**	0.93	**0.94**
	Person	0.95	**0.96**	0.95	—
En	**Number**	0.95	**0.96**	**0.96**	**0.96**
	Person	0.98	**0.99**	0.98	**0.99**
Fr	**Gender**	0.92	0.92	0.92	**0.93**
	Number	**0.94**	0.93	**0.94**	**0.94**
	Person	0.96	**0.97**	0.96	**0.97**
Ru	**Case**	0.98	**0.99**	0.98	**0.99**
	Gender	0.96	0.97	0.96	**0.98**
	Number	0.98	0.98	0.98	**0.99**
	Person	0.98	**0.99**	0.98	**0.99**

Table 6: The results of the probing classifier on **Morphosyntactic Features** tasks for pre-trained models. The scores are averaged across all layers. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Lang	Probing Task	D-BERT	MiniLM	BERT	XLM-R
De	**Case**	0.91	0.92	0.91	**0.93**
	Gender	0.91	0.91	0.92	**0.93**
	Number	0.94	0.94	0.94	**0.95**
	Person	0.95	0.96	0.96	**0.97**
En	**Number**	0.96	0.96	0.96	**0.97**
	Person	0.98	**0.99**	0.98	**0.99**
Fr	**Gender**	**0.93**	0.92	**0.93**	**0.93**
	Number	0.94	0.93	**0.95**	0.94
	Person	0.96	**0.97**	0.96	**0.97**
Ru	**Case**	**0.99**	**0.99**	**0.99**	**0.99**
	Gender	0.97	0.97	0.97	**0.98**
	Number	**0.99**	**0.99**	**0.99**	**0.99**
	Person	**0.99**	**0.99**	**0.99**	**0.99**

Table 7: The results of the probing classifier on **Morphosyntactic Features** tasks for fine-tuned models. The scores are averaged across all layers. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Lang	Probing Task	D-BERT	MiniLM	BERT	XLM-R
De	**Gender**	—	—	—	—
	Number	—	—	—	—
En	**Gender**	0.52	0.50	**0.53**	0.51
	Number	0.66	0.59	**0.68**	0.67
Fr	**Gender**	**0.72**	0.68	0.66	0.69
	Number	0.70	0.65	0.71	**0.73**
Ru	**Case**	0.67	0.61	0.74	**0.78**
	Gender	0.68	0.67	0.67	**0.73**
	Number	0.67	0.63	0.71	**0.75**
	Person	0.62	0.51	0.59	**0.69**

Table 8: The results of the probing classifier on **Masked Tokens** tasks for pre-trained models. The scores are averaged across all layers. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Lang	Probing Task	D-BERT	MiniLM	BERT	XLM-R
De	**Gender**	—	—	—	—
	Number	—	—	—	—
En	**Gender**	0.51	0.51	0.51	**0.52**
	Number	0.64	0.61	0.64	**0.66**
Fr	**Gender**	0.67	0.60	**0.69**	0.62
	Number	0.63	0.65	**0.69**	0.59
Ru	**Case**	0.67	0.62	0.67	**0.70**
	Gender	0.67	0.62	0.50	**0.68**
	Number	**0.66**	0.60	0.5	0.63
	Person	0.52	0.52	0.53	**0.55**

Table 9: The results of the probing classifier on **Masked Tokens** tasks for fine-tuned models. The scores are averaged across all layers. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Lang	Probing Task	D-BERT	MiniLM	BERT	XLM-R
De	**Article Removal**	**0.97**	0.96	0.96	**0.97**
	Deixis Word Number	0.65	0.63	0.66	**0.73**
	Subject Number	0.6	0.66	0.68	**0.72**
	Predicate Number	0.67	0.67	0.72	**0.77**
En	**Article Removal**	**0.98**	0.97	0.97	0.97
	Stop-words Removal	**0.99**	**0.99**	0.98	**0.99**
	Subject Number	**0.53**	**0.53**	0.51	0.51
	Predicate Number	0.51	0.53	0.52	**0.59**
Fr	**Article Removal**	0.96	0.96	0.95	**0.97**
	Subject Number	0.63	0.65	**0.71**	**0.71**
	Predicate number	0.67	0.71	0.74	**0.76**
Ru	**Stop-words Removal**	0.95	**0.96**	0.94	**0.96**
	Subject Case	0.72	0.75	0.77	**0.82**
	Subject Number	0.63	0.68	0.7	**0.76**
	Predicate Gender	0.63	0.64	0.67	**0.71**
	Predicate Number	0.64	0.67	0.71	**0.75**
	Predicate Person	0.77	0.8	0.81	**0.86**

Table 10: The results of the probing classifier on **Perturbations** tasks for pre-trained models. The scores are averaged across all layers. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

Lang	Probing Task	D-BERT	MiniLM	BERT	XLM-R
De	**Article Removal**	**0.98**	**0.98**	**0.98**	**0.98**
	Deixis Word Number	0.63	0.63	0.69	**0.72**
	Subject Number	0.62	0.62	0.68	**0.71**
	Predicate Number	0.67	0.64	0.7	**0.75**
En	**Article Removal**	**0.99**	0.98	**0.99**	**0.99**
	Stop-words Removal	**0.99**	**0.99**	**0.99**	**0.99**
	Subject Number	0.54	0.52	**0.57**	0.55
	Predicate Number	0.51	0.52	0.54	**0.6**
Fr	**Article**	0.97	0.96	0.96	**0.98**
	Subject Number	0.65	0.67	0.73	**0.76**
	Predicate Number	0.67	0.64	0.72	**0.76**
Ru	**Stop-words Removal**	0.96	0.95	0.96	**0.97**
	Subject Case	0.75	0.75	0.79	**0.84**
	Subject Number	0.65	0.65	0.72	**0.77**
	Predicate Gender	0.63	0.62	0.67	**0.71**
	Predicate Number	0.62	0.62	0.7	**0.75**
	Predicate Person	0.79	0.8	0.8	**0.85**

Table 11: The results of the probing classifier on **Perturbations** tasks for fine-tuned models. The scores are averaged across all layers. Languages: **Ru**=Russian, **En**=English, **De**=German, **Fr**=French.

E Individual Neuron Analysis

Property-wise analysis Figures 4 – 7 depict property-wise neuron distribution for French, Russian, German and English.

Layer-wise analysis Figures 8 – 10 demonstrate the results of the individual neuron analysis on **Subject Number** perturbation task for Russian, English and French.

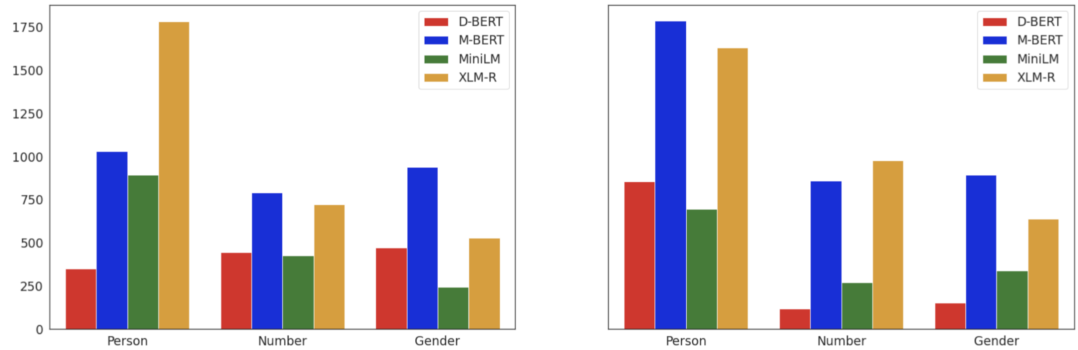

Figure 4: Number of neurons per each property for French. Y-axis=Number of neurons. Left: pre-trained models. Right: fine-tuned models.

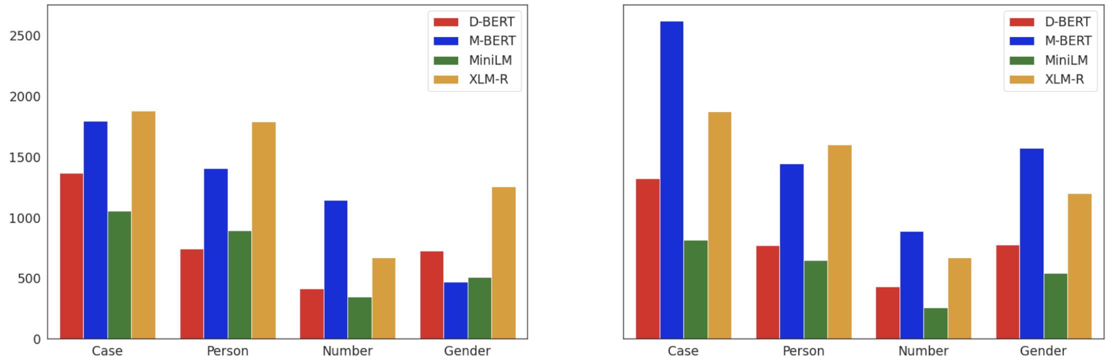

Figure 5: Number of neurons per each property for Russian. Y-axis=Number of neurons. Left: pre-trained models. Right: fine-tuned models.

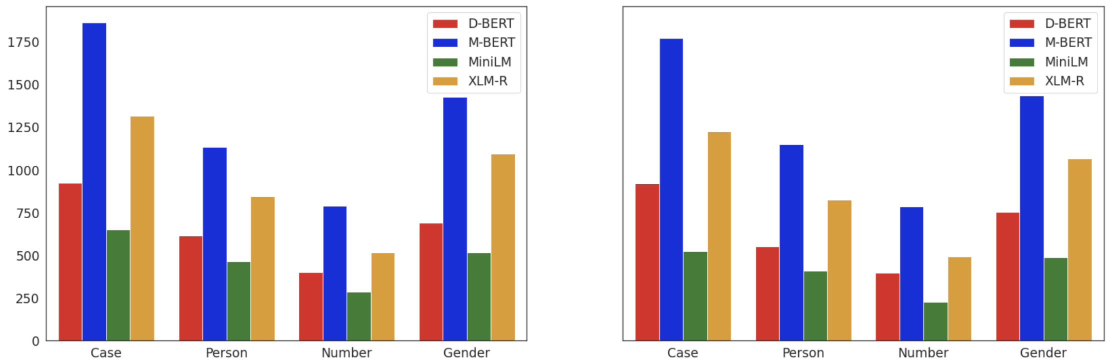

Figure 6: Number of neurons per each property for German. Y-axis=Number of neurons. Left: pre-trained models. Right: fine-tuned models.

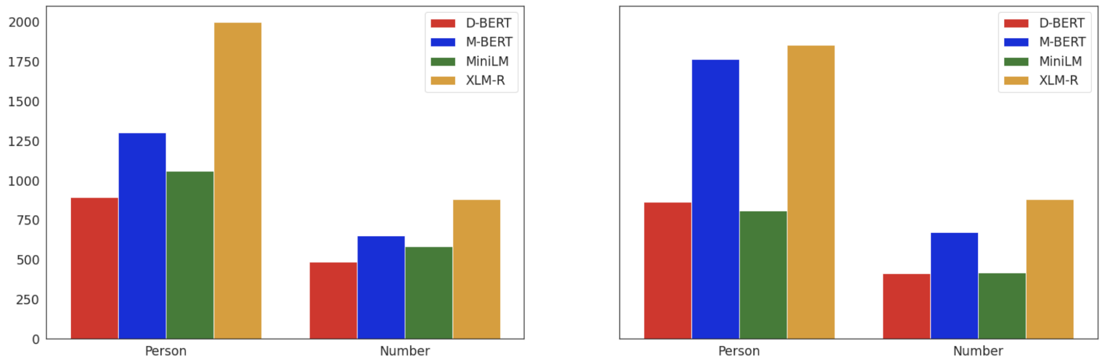

Figure 7: Number of neurons per each property for English. Y-axis=Number of neurons. Left: pre-trained models. Right: fine-tuned models.

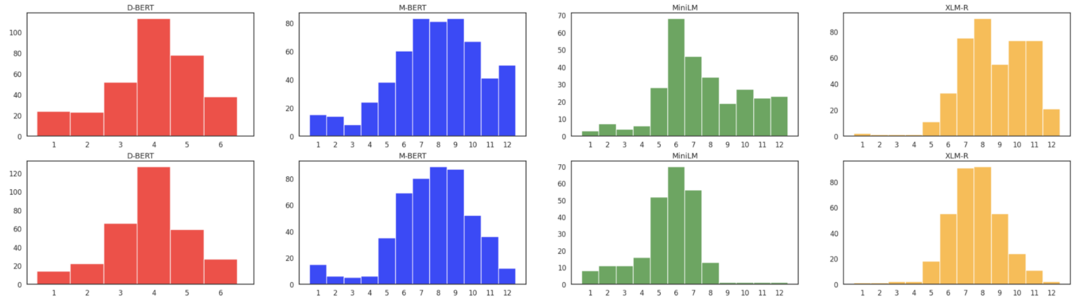

Figure 8: The distribution of top neurons over **Subject Number** perturbation task for each model (**Russian**). X-axis=Layer index number. Y-axis=Number of neurons. Top: pre-trained models. Bottom: fine-tuned models.

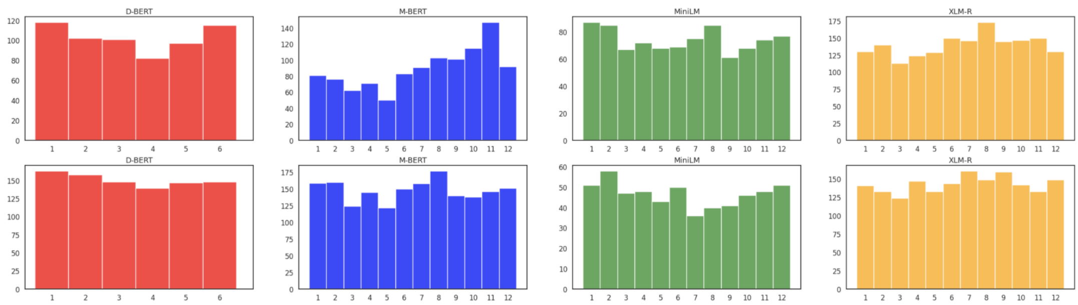

Figure 9: The distribution of top neurons over **Subject Number** perturbation task for each model (**English**). X-axis=Layer index number. Y-axis=Number of neurons. Top: pre-trained models. Bottom: fine-tuned models.

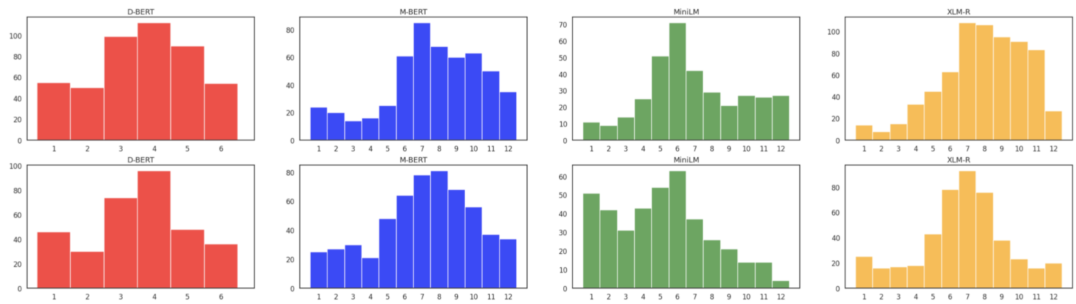

Figure 10: The distribution of top neurons over **Subject Number** perturbation task for each model (**French**). X-axis=Layer index number. Y-axis=Number of neurons. Top: pre-trained models. Bottom: fine-tuned models.

F POS-Tagging Performance

Tables 12 – 15 describe the results of the fine-tuning on POS-tagging task for each language.

Model / Metric	Accuracy	F1	Precision	Recall
M-BERT	**0.98**	**0.98**	**0.98**	**0.98**
DistilBERT	**0.98**	**0.98**	**0.98**	**0.98**
MiniLM	**0.98**	**0.98**	**0.98**	**0.98**
XLM-R	**0.98**	**0.98**	**0.98**	**0.98**

Table 12: Metrics of the models fine-tuned for POS-tagging task for German.

Model / Metric	Accuracy	F1	Precision	Recall
M-BERT	**0.96**	0.95	0.95	0.95
DistilBERT	**0.95**	0.94	0.94	0.94
MiniLM	**0.95**	0.94	0.94	0.94
XLM-R	**0.96**	**0.96**	**0.96**	**0.96**

Table 13: Metrics of the models fine-tuned for POS-tagging task for English.

Model / Metric	Accuracy	F1	Precision	Recall
M-BERT	**0.98**	0.97	0.97	0.97
DistilBERT	**0.97**	**0.97**	**0.97**	**0.97**
MiniLM	**0.97**	0.96	0.96	0.96
XLM-R	**0.98**	**0.98**	**0.98**	**0.98**

Table 14: Metrics of the models fine-tuned for POS-tagging task for French.

Model / Metric	Accuracy	F1	Precision	Recall
M-BERT	**0.99**	**0.99**	**0.99**	**0.99**
DistilBERT	**0.99**	**0.99**	**0.99**	**0.99**
MiniLM	**0.99**	0.98	0.98	0.98
XLM-R	**0.99**	**0.99**	**0.99**	**0.99**

Table 15: Metrics of the models fine-tuned for POS-tagging task for Russian.