Probing for Predicate Argument Structures in Pretrained Language Models

Simone Conia[1] and **Roberto Navigli**[2]
Sapienza NLP Group
[1]Department of Computer Science
[2]Department of Computer, Control and Management Engineering
Sapienza University of Rome
conia@di.uniroma1.it navigli@diag.uniroma1.it

Abstract

Thanks to the effectiveness and wide availability of modern pretrained language models (PLMs), recently proposed approaches have achieved remarkable results in dependency- and span-based, multilingual and cross-lingual Semantic Role Labeling (SRL). These results have prompted researchers to investigate the inner workings of modern PLMs with the aim of understanding how, where, and to what extent they encode information about SRL. In this paper, we follow this line of research and probe for predicate argument structures in PLMs. Our study shows that PLMs do encode semantic structures directly into the contextualized representation of a predicate, and also provides insights into the correlation between predicate senses and their structures, the degree of transferability between nominal and verbal structures, and how such structures are encoded across languages. Finally, we look at the practical implications of such insights and demonstrate the benefits of embedding predicate argument structure information into an SRL model.

1 Introduction

Semantic Role Labeling (SRL) is often defined informally as the task of automatically answering the question "*Who did What to Whom, Where, When and How?*" (Màrquez et al., 2008) and is, therefore, thought to be a fundamental step towards Natural Language Understanding (Navigli, 2018). Over the past few years, SRL has started to gain renewed traction, thanks mainly to the effectiveness and wide availability of modern pretrained language models (PLMs), such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and BART (Lewis et al., 2020). Current approaches have, indeed, attained impressive results on standard evaluation benchmarks for dependency- and span-based, multilingual and cross-lingual SRL (He et al., 2019; Li et al., 2019; Cai and Lapata, 2020; Conia and Navigli, 2020; Blloshmi et al., 2021; Conia et al., 2021).

Despite the remarkable benefits provided by the rich contextualized word representations coming from PLMs, the novelties introduced in recent state-of-the-art models for SRL revolve primarily around developing complexities on top of such word representations, rather than investigating what happens inside a PLM. For example, the SRL systems of He et al. (2019) and Conia and Navigli (2020) take advantage only of BERT's uppermost hidden layers to build their input word representations. However, the revolution that PLMs have sparked in numerous areas of Natural Language Processing (NLP) has motivated researchers in the community to investigate the inner workings of such models, with the aim of understanding how, where, and to what extent they encode information about specific tasks. This research has revealed that different layers encode significantly different features (Tenney et al., 2019; Vulić et al., 2020). In perhaps one of the most notable studies in this direction, Tenney et al. (2019) demonstrated empirically that BERT "rediscovers" the classical NLP pipeline, highlighting that the lower layers tend to encode mostly lexical-level information while upper layers seem to favor sentence-level information.

Although recent analyses have already provided important insights into which layers of a PLM are more relevant for SRL and how their relative importance is affected by the linguistic formalism of choice (Kuznetsov and Gurevych, 2020), not only do these analyses treat SRL as an atomic task but they also do not explore taking advantage of their insights to improve current state-of-the-art SRL systems. Indeed, the SRL pipeline is usually divided into four main steps: predicate identification and disambiguation, and argument identification and classification. To address this gap, in this paper we therefore take an in-depth look at how predicate senses and their predicate argument struc-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4622 - 4632

tures (PASs) are encoded across different layers of different PLMs. On the one hand, we provide new insights into the capability of these models to capture complex linguistic features, while on the other, we show the benefits of embedding such features into SRL systems to improve their performance.

Our contributions can be summarized as follows:

- We probe PLMs for PASs: do PLMs encode the argument structure of a predicate in its contextual representation?
- We show that, even though a PAS is defined according to a predicate sense, senses and argument structures are encoded at different layers in PLMs;
- We demonstrate empirically that verbal and nominal PASs are represented differently across the layers of a PLM;
- Current SRL systems do not discriminate between nominal and verbal PASs: we demonstrate that, although there exists some degree of transferability between the two, an SRL system benefits from treating them separately;
- We find that PAS information is encoded similarly across two very different languages, English and Chinese, in multilingual PLMs;
- We corroborate our findings by proposing a simple approach for integrating predicate-argument structure knowledge into an SRL architecture, attaining improved results on standard gold benchmarks.

We hope that our work will contribute both to the understanding of the inner workings of modern pretrained language models and to the development of more effective SRL systems. We release our software for research purposes at https://github.com/SapienzaNLP/srl-pas-probing.

2 Related Work

Probing pretrained language models. The unprecedented capability of modern PLMs to provide rich contextualized input representations took the NLP community by storm. Alongside the rising wave of successes collected by PLMs in an ever increasing number of areas, researchers started to question and investigate what happens inside these models and what they really capture, probing for knowledge and linguistic properties (Hewitt and Manning, 2019; Chi et al., 2020; Vulić et al., 2020). This body of work quickly attracted increasing attention and grew to become a field of study with a name of its own: BERTology (Rogers et al., 2020). Probing a PLM usually consists in defining a very precise task (e.g., identifying whether two words are linked by a syntactic or semantic relation), and then in designing and training a simple model, called a *probe*, to solve the task using the contextualized representations provided by the PLM. The idea is to design a probe that is as simple as possible, often consisting of a single-layer model: if the probe is able to address the task, then it must be thanks to the contextual information captured by the PLM as the expressiveness of the probe itself is limited by its simplicity. One could argue that some complex relations may require a non-linear probe (White et al., 2021) which can reveal hidden information as long as it is accompanied by control experiments (Hewitt and Liang, 2019) to verify that it is still extracting information from the underlying PLM, rather than merely learning to solve the probing task. Over the past few years, these probing techniques have been used to great effect and revealed that PLMs have been "rediscovering" the classical NLP pipeline (Tenney et al., 2019), and that they often encode distances between syntactic constituents (Hewitt and Liang, 2019), lexical relations (Vulić et al., 2020) and morphology (Chi et al., 2020), *inter alia*.

Probing techniques for SRL. As in several other fields of NLP, recent studies have aimed to shed some light on how, where and to what extent PLMs encode information relevant to SRL. Among others, Tenney et al. (2019) devised an edge probing mechanism aimed at ascertaining the capability of BERT to identify which semantic role ties a given predicate to a given argument span, and showed that this task is "solved" mainly by the middle layers of BERT. Toshniwal et al. (2020) proposed and compared several techniques for better combining the contextualized representations of a PLM, finding that applying max pooling or performing a weighted average are two robust strategies for SRL. More recently, Kuznetsov and Gurevych (2020) designed a probe to analyze how different linguistic ontologies – essential to the task in that they define predicate senses and semantic roles explicitly – require features that are encoded at different layers of a PLM. In this paper, we follow the line of research laid out by the afore-

mentioned work, probing PLMs with the objective of understanding where and to what extent they encode a predicate argument structure into the contextualized representation of a predicate.

Recent advances in SRL. Thanks to their effectiveness, PLMs are now the *de facto* input representation method in SRL (He et al., 2019; Li et al., 2019; Conia and Navigli, 2020; Blloshmi et al., 2021). Recently proposed approaches have achieved impressive results on several gold benchmarks (Hajič et al., 2009; Pradhan et al., 2012), both in span-based and in dependency-based SRL, but also in multilingual and cross-lingual SRL, even though there still seems to be a significant margin for improvement in out-of-domain settings. The innovations put forward by such approaches, however, have mainly focused on architectural novelties built on top of PLMs: Cai et al. (2018) proposed the first end-to-end architecture; He et al. (2019) and Cai and Lapata (2019) successfully exploited syntax in multilingual SRL; Marcheggiani and Titov (2020) took advantage of GCNs to capture distant semantic relations; Conia and Navigli (2020) devised a language-agnostic approach to bridge the gap in multilingual SRL; Blloshmi et al. (2021) and Paolini et al. (2021) tackled the task as a sequence generation problem; Conia et al. (2021) introduced a model to perform cross-lingual SRL across heterogeneous linguistic inventories. However, if we look back at past work, it is easy to realize that we lack a study that provides an in-depth look into PLMs and a hint at how to better exploit them in future SRL systems.

3 Probing for Predicate Senses and Their Predicate-Argument Structures

As mentioned above, some studies have already investigated how semantic knowledge is distributed among the inner layers of current PLMs, finding that information useful for SRL is mainly stored in their middle layers (Tenney et al., 2019). However, such studies have considered SRL as an atomic task, while instead the SRL pipeline can be thought of as being composed of four different subtasks:

1. **Predicate identification**, which consists in identifying all those words or multi-word expressions that denote an action or an event in the input sentence;
2. **Predicate sense disambiguation**, which requires choosing the most appropriate *sense* or *frame* for each predicate identified, as the same predicate may denote different meanings or define different semantic scenarios depending on the context;
3. **Argument identification**, which consists in selecting the parts of the input text that are "semantically" linked as arguments to an identified and disambiguated predicate;
4. **Argument classification**, which is the task of determining which kind of semantic relation, i.e., semantic role, governs each predicate-argument pair.

For our study, it is important to note that, in many popular ontologies for SRL, predicate senses or frames are often tightly coupled to their possible semantic roles. In other words, the set of possible semantic roles that can be linked to a predicate p is defined according to the sense or frame of p. Hereafter, given a predicate p, we refer to its set of possible semantic roles as the *roleset* of p. For example, the predicate love as in "He loved everything about her" belongs to the FrameNet (Baker et al., 1998) frame *experiencer_focused_emotion* which defines a roleset composed of {Experiencer, Content, . . ., Degree}. The same predicate sense has different rolesets in other ontologies, for example {ARG0 (lover), ARG1 (loved)} in the English PropBank (Palmer et al., 2005) and {Experiencer, Stimulus, . . ., Cause} in VerbAtlas (Di Fabio et al., 2019).

3.1 Predicate Senses and Their Rolesets

Since rolesets are often defined according to predicate senses, it is interesting to investigate whether current pretrained language models store important features about senses and rolesets in their hidden layers. To this end, we formulate two simple probing tasks:

- **Sense probing**, which consists in predicting the sense s of a predicate p from the contextual vector representation $\mathbf{x}_p$ of p, where $\mathbf{x}_p$ is obtained from a pretrained language model.
- **Roleset probing**, which consists in predicting the semantic roles $\{r_1, r_2, \ldots, r_n\}$ that appear linked to a predicate p from its contextual representation $\mathbf{x}_p$, where $\mathbf{x}_p$ is obtained from a pretrained language model.

For the choice of $\mathbf{x}_p$, we compare four different options:

- **Random:** initializing the weights of the language model at random provides a simple control baseline to attest the ability of a probe to "learn the probing task", i.e. learning to associate random inputs to correct labels;

- **Static:** $\mathbf{x}_p$ is the input embedding of the pretrained language model corresponding to p, e.g., the non-contextual representation before the Transformer layers in BERT.[1]

- **Top-4:** $\mathbf{x}_p$ is the concatenation of the topmost four hidden layers of the language model: this is the configuration used in some of the recently proposed approaches for full SRL systems (Conia and Navigli, 2020);

- **W-Avg:** $\mathbf{x}_p$ is the weighted average of all the hidden layers of the language model, where the weights for each layer are learned during training (the larger the weight the more important its corresponding layer is for the probing task).

For each probing task, we train[2] two simple probes, a linear classifier and a non-linear[3] classifier, on the verbal predicate instances of the English training datasets provided as part of the CoNLL-2009 shared task for dependency-based SRL (Hajič et al., 2009).

3.2 Probing Results

Results on sense probing. Table 1 reports the results of our linear and non-linear probes on predicate sense disambiguation when using different types of input representations $\mathbf{x}_p$, namely, Static, Random, Last-4 and W-Avg, of an input predicate p in context. The Random baseline is able to disambiguate well (84.8% in Accuracy using BERT-base-cased), which is, however, unsurprising since CoNLL-2009 is tagged with PropBank labels and most of the predicates are annotated with their first sense (e.g., *buy.01*, *sell.01*). Interestingly, static representations from all four language models do not contain much more information about predicate senses than random representations. Using the topmost four hidden layers, instead, provides a substantial improvement over static representations for all language models (e.g., +6% in Accuracy for BERT-base-cased), lending credibility to the fact that context is key for the disambiguation process. Most notably, the best representation for the sense probing task is consistently obtained by performing a weighted average of all the hidden layers of the language model. This shows that important predicate sense information is not stored only in the topmost hidden layers and, therefore, also hints at the possibility that state-of-the-art architectures, such as those of He et al. (2019) and Conia and Navigli (2020), do not exploit pretrained language models to their fullest. Finally, it is interesting to note that linear and non-linear probes obtain similar results, showing that sense-related information can easily be extracted without the need for a complex probe.

		BERT	RoBERTa	m-BERT	XLM-R
Linear	Random	84.8	85.6	–	–
	Static	84.7	86.6	–	–
	Top-4	92.8	93.4	–	–
	W-Avg	**94.4**	**94.5**	–	–
Non-Linear	Random	84.3	83.6	83.7	84.2
	Static	86.4	86.6	86.1	86.1
	Top-4	93.2	93.6	92.3	93.3
	W-Avg	**94.2**	**94.8**	**93.4**	**94.2**

Table 1: Results on **sense probing** in terms of Accuracy (%) for the Random, Static, Top-4 and W-Avg probes using different pretrained language models, namely, BERT (base-cased), RoBERTa (base), multilingual BERT (base) and XLM-RoBERTa (base). Using a weighted average of all the hidden layers is a better choice than using the concatenation of the topmost four layers as in Conia and Navigli (2020).

Results on roleset probing. Table 2 reports the results on roleset identification obtained by our linear and non-linear probes when using different types of input representations $\mathbf{x}_p$, namely, Static, Random, Top-4 and W-Avg, of an input predicate p in context. For this task, we measure the performance of a probe in terms of micro-averaged F1 score, taking into account partially correct predictions, e.g., the system is partially rewarded for predicting {ARG0, ARG1} instead of {ARG0, ARG2}. As is the case for sense probing, our simple Random baseline is able to identify the correct

[1] In case of a predicate composed of multiple subtokens, $\mathbf{x}_p$ is the average of the vector representations of its subtokens.

[2] We train each probe for 20 epochs using Adam (Kingma and Ba, 2015) as the optimizer with a learning rate of 1e-3. As is customary in probing studies, the weights of the pretrained language models are kept frozen during training. We use the pretrained language models made available by Huggingface's Transformers library (Wolf et al., 2020).

[3] We use the Swish activation function (Ramachandran et al., 2018) for our non-linear probes.

roleset for a predicate in context with a satisfactory performance (72.8% in F1 score using BERT-base-cased). Indeed, most predicates have at least one argument tagged with either ARG0 or ARG1, which in PropBank usually correspond to agentive and patientive proto-roles, respectively; we hypothesize that the Random probe merely learns to bias its predictions towards these very common semantic roles. Differently from in the sense probing task, the non-linear probe seems to perform better and achieve higher scores than the linear one. However, this does not mean that roleset-related features are "stored" non-linearly in PLMs. Indeed, one can notice that the random non-linear probe also performs better than its linear counterpart, suggesting that the higher score is due to the greater expressiveness of the probe, which "learns" the task rather than "extracting" information from the underlying PLM, i.e., the *selectivity* (Hewitt and Liang, 2019) of a non-linear probe is not greater than that of a linear probe in this task.

Despite the fact that the roleset probing task is more difficult than the sense probing one, we can observe a similar trend in the results: the Top-4 probe is substantially better than the Static probe, but W-Avg consistently outperforms Top-4, strongly suggesting that future approaches will need to use all the layers to take full advantage of the knowledge encoded within PLMs. We stress that not exploiting all the inner layers of a PLM is an illogical choice, since the cost of computing a weighted average of their hidden representations is negligible compared to the overall computational cost of a Transformer-based architecture.

On the correlation between senses and rolesets. Thus far, we have seen empirical evidence that PLMs encode important features about predicate senses and their rolesets across all their hidden layers, not just the topmost ones often used in the literature by current models for SRL. However, one may wonder how such features are distributed across these hidden layers. As we have already discussed above, predicate senses and their rolesets are tightly coupled: do PLMs distribute sense and roleset features similarly over their inner layers?

To answer this question, we resort to the W-Avg probe we introduced above. Indeed, its peculiarity is that it learns to assign a different weight to each hidden layer of a PLM: in order to minimize the training loss, the W-Avg probe will assign a larger weight to those layers that are most beneficial, i.e., to those layers that express features that are more relevant for the probing task. Therefore, we extract such layer weights learned by our probes for the two tasks we are studying – predicate sense disambiguation and roleset identification – and compare these learned weights, as shown in Figure 1 (top, blue charts). Interestingly, and perhaps surprisingly, the W-Avg probe learns a different weight distribution for the two probing tasks, even though rolesets are often defined on the basis of predicate senses in many popular ontologies for SRL. We can observe that predicate sense features are encoded more uniformly across the hidden layers of BERT or, equivalently, that the probe assigns similar weights to each hidden layer, slightly preferring the topmost ones (Figure 1, top-left). However, this is not the case for the roleset probing task, in which the probe mostly relies on the hidden layers going from the 6th to the 10th, almost disregarding the bottom and top ones. Furthermore, we can observe the same negative correlation within the distributions of the layer weights learned for senses and rolesets when using RoBERTa, albeit the divergence is slightly less accentuated (Figure 1, top-right).

		BERT	RoBERTa	m-BERT	XLM-R
Linear	Random	72.8	72.8	–	–
	Static	75.1	75.3	–	–
	Top-4	85.3	85.3	–	–
	W-Avg	**85.7**	**86.1**	–	–
Non-Linear	Random	75.9	75.9	75.8	75.7
	Static	76.3	76.5	76.2	76.3
	Top-4	89.2	88.8	88.0	88.9
	W-Avg	**89.4**	**89.3**	**88.8**	**89.1**

Table 2: Results on **roleset probing** in terms of F1 Score (%) for the Random, Static, Top-4 and W-Avg probes using different pretrained language models, namely, BERT (base-cased), RoBERTa (base), multilingual BERT (base) and XLM-RoBERTa (base). As for the sense probing task, using the a weighted average of all the hidden layers provides richer features to the probes.

3.3 Verbal and Nominal Predicates

One aspect that is often overlooked when designing and proposing novel architectures for SRL is that not all predicates are verbs. In English, it is easy to find examples of nouns that evoke or imply a predication, such as *producer*, *driver*, and *writer*. Most common nominal predicates are "verb-derived" or "deverbal" as their roleset is derived from their corresponding verbal predicates. This is why, per-

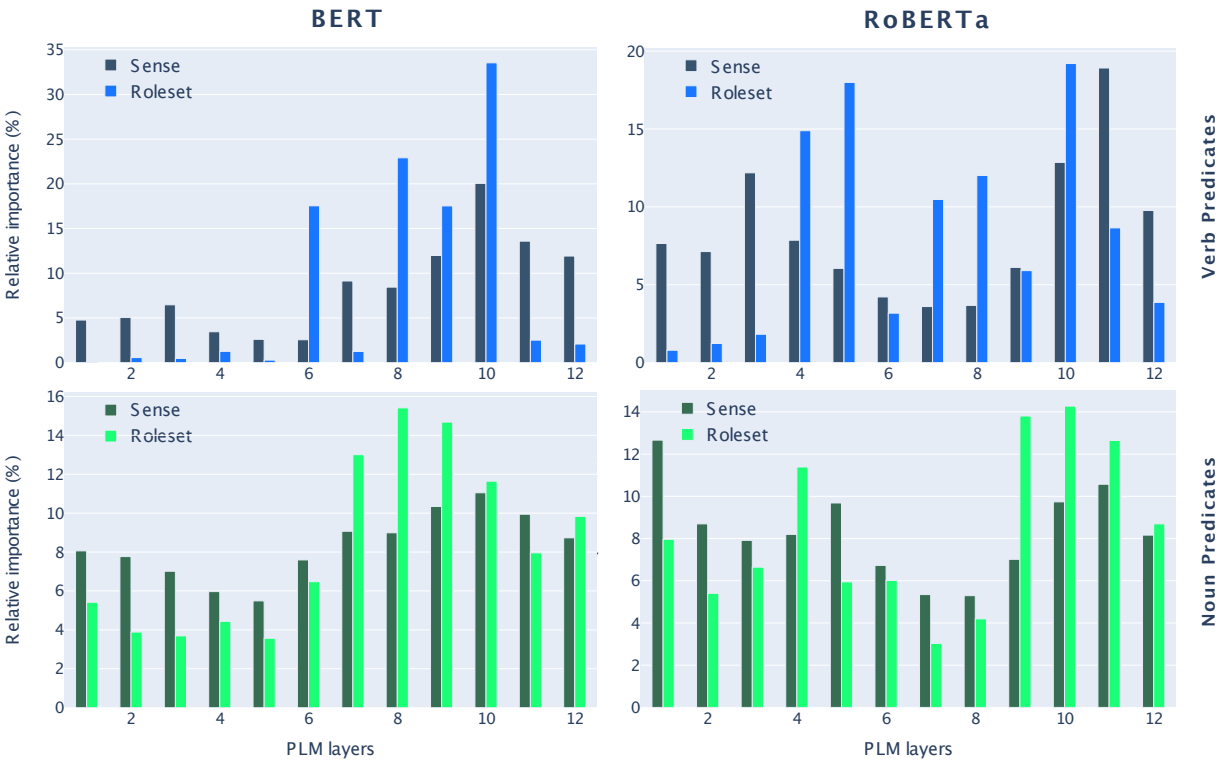

Figure 1: Relative importance (%) of each layer of BERT (left) and RoBERTa (right) for sense probing and roleset probing. Verbal predicates (top, blue): the most important layers of a PLM for roleset probing are the middle layers, especially for BERT, in which the top and the bottom layers are almost completely discarded. Nominal predicates (bottom, green): the importance of each layer follows the same trend for both sense and roleset probing.

PLM	Trained on	Verbs (F1)	Nouns (F1)
Random	Verbs	72.0	–
Random	Nouns	–	68.5
BERT	Verbs	85.7	63.3
BERT	Nouns	67.5	77.5
RoBERTa	Verbs	86.1	64.7
RoBERTa	Nouns	67.5	78.3

Table 3: Results in terms of F1 score (%) on zero-shot roleset identification when a probe is trained on verbal predicates and evaluated on nominal predicates, and vice versa. Interestingly, a probe trained on verbal predicates performs worse than a random probe on nominal predicates, demonstrating that knowledge transfer between predicate types is not trivial.

haps, current state-of-the-art approaches do not distinguish between verbal and nominal predicates.[4] However, nominal predicates also possess peculiarities that do not appear in their verbal counterparts; for example, a nominal predicate can be its own argument, e.g., *writer* is the agent itself of the action *write*.

We take this opportunity to investigate how nominal predicate senses and their rolesets are encoded by PLMs in their inner layers. We train a W-Avg probe on the sense and roleset probing tasks, focusing only on the nominal predicate instances in CoNLL-2009. Figure 1 (bottom, green charts) shows the weights learned for the sense and roleset probing tasks when using BERT (bottom-left) and RoBERTa (bottom-right): we can immediately observe that, differently from verbal predicates, the weight distributions learned for nominal senses and their rolesets follow the same trend in both PLMs. In other words, despite the fact that most nominal predicates are verb-derived, their information is encoded dissimilarly and distributed across different layers compared to those of verbal predicates.

We confirm our hunch by evaluating the ability of a W-Avg probe trained on roleset identification for verbal predicates only to also perform roleset identification for nominal predicates in a zero-shot fashion, and vice versa. Although, from a first glance at the results reported in Table 3, our simple model seems to be able to perform nominal roleset identification after being trained only on verbal

[4]We note that, in general, languages – English included – also possess, sometimes in extensive quantities, predicates that are neither verbal nor nominal. For example, Japanese prominently features adjectival predicates.

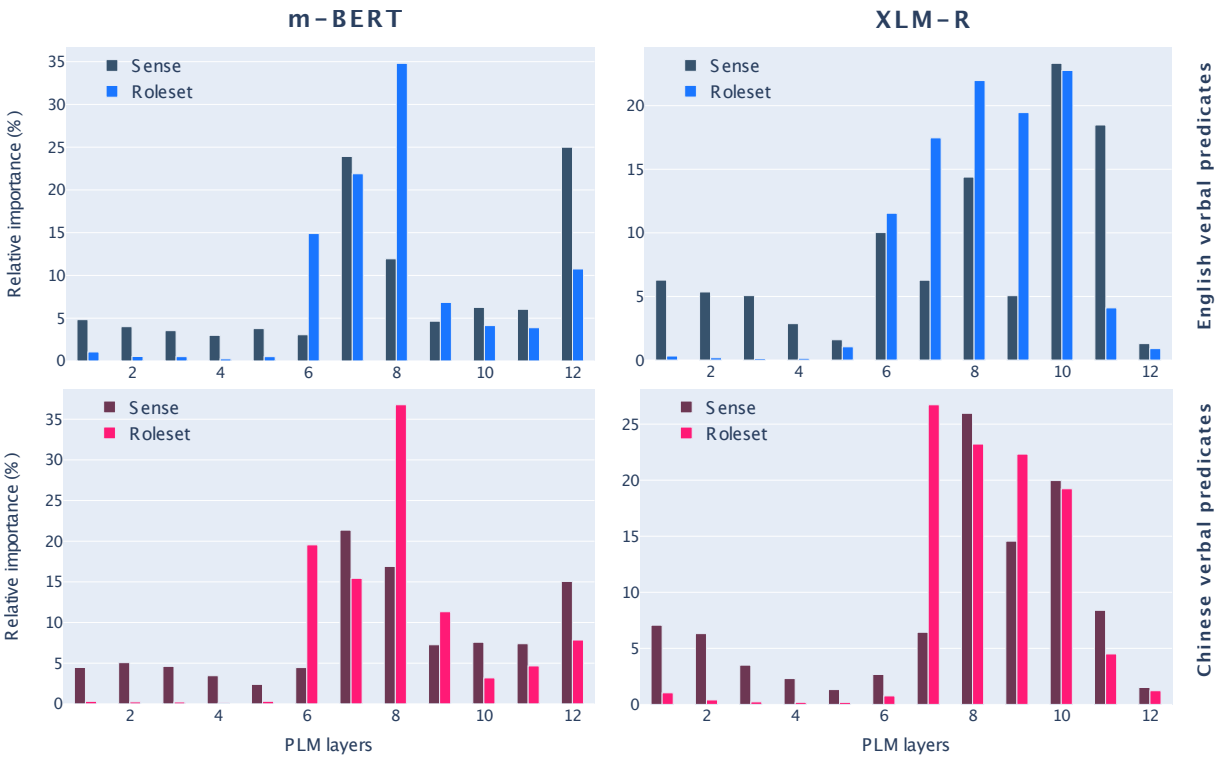

Figure 2: Relative importance (%) of each hidden layer of multilingual BERT (left) and XLM-RoBERTa (right) for sense probing and roleset probing. Results in English are in blue (top), whereas results in Chinese are in red (bottom).

rolesets, the performance is actually worse than a control probe, which is trained with a randomly initialized model on nominal roleset identification. In general, our analysis provides an empirical explanation for why recent approaches for nominal SRL adapted from verbal SRL are still struggling to learn general features across different predicate types, despite initial promising results (Klein et al., 2020; Zhao and Titov, 2020).

3.4 Senses and Rolesets Across Languages

We conclude our analysis on predicate senses and their rolesets with another important finding: multilingual PLMs encode both predicate sense and roleset information at similar layers across two very different languages, English and Chinese. In order to support this statement, we train an W-Avg probe on both sense disambiguation and roleset identification, first on the English verbal predicates from the training split of CoNLL-2009 and then on the Chinese verbal predicates from the training split of CoNLL-2009.

Figure 2 shows the distributions of the learned weights for each hidden layer of two language models, multilingual BERT (left) and XLM-RoBERTa (right). In particular, we observe that the probe learns to almost completely discard the first five layers of multilingual BERT for roleset identification in both English (top-left) and Chinese (bottom-left), while assigning similar weights across English and Chinese to the other hidden layers, with the 8th layer being relatively important in both languages. Overall, Figure 2 supports the evidence that both multilingual BERT and XLM-RoBERTa encode the same type of "semantic knowledge" at roughly the same hidden layers across languages, supporting the findings by Conneau et al. (2020) and indicating a possible direction for future work in cross-lingual transfer learning for SRL.

4 Integrating Predicate-Argument Structure Knowledge

Now that we have provided an in-depth look at how sense and roleset information is encoded at different inner layers of current PLMs (Section 3.2), highlighted the differences in how PLMs encode verbal and nominal predicates (Section 3.3), and revealed that multilingual PLMs capture semantic knowledge at similar layers across two diverse languages (Section 3.4), one may wonder how we can take advantage in a practical setting of what we have learned so far. In this Section, we study how

we can improve a modern system for end-to-end SRL by integrating sense and roleset knowledge into its architecture.

4.1 Model Description

In what follows, we briefly describe the architecture of our baseline model, which is based on that proposed by Conia and Navigli (2020). Notice that, even though we refer to this model as our baseline, its end-to-end architecture rivals current state-of-the-art approaches, such as Blloshmi et al. (2021), Conia et al. (2021) and Paolini et al. (2021).

Given an input sentence $\mathbf{w}$, the model computes a contextual representation $\mathbf{x}_i$ for each word w_i in $\mathbf{w}$ by concatenating the representations obtained from the four topmost layers of a pretrained language model. These contextual word representations are then processed by a stack of "fully connected" BiLSTM layers in which the input to the i-th BiLSTM layer is the concatenation of the inputs of all previous BiLSTM layers in the stack, obtaining a sequence $\mathbf{h}$ of refined encodings. These encodings $\mathbf{h}$ are made "predicate-aware" by concatenating each $\mathbf{h}_i$ of w_i to the representation $\mathbf{h}_p$ of each predicate p in the sentence, and finally processed by another stack of fully-connected BiLSTMs, resulting in a sequence $\mathbf{a}$ of argument encodings. We refer to Conia and Navigli (2020) for further details about the architecture of our baseline model.

Enhancing the SRL model. Based on our observations and analyses in the Sections above, we put forward three simple enhancements to our strong baseline model:

- Representing words using a weighted average of all the inner layers of the underlying language model, since we now know that semantic features important for the task are scattered across all the layers of a PLM;
- Using two different sets of weights to compute different weighted averages for predicate senses and predicate arguments, as semantic features important for the two tasks are distributed differently across the inner layers of the underlying PLM;
- Adding a secondary task to predict rolesets from a predicate representation $\mathbf{h}_p$ in a multi-task learning fashion.

	P	R	F1
$\text{BERT}_{\text{base}}$ – baseline	91.8	91.9	91.8
$\text{BERT}_{\text{base}}$ – W-Avg	91.9	92.0	91.9
$\text{BERT}_{\text{base}}$ – 2×W-Avg	92.1	92.1	92.1
$\text{BERT}_{\text{base}}$ – 2×W-Avg + MT	92.2	92.2	92.2
$\text{BERT}_{\text{large}}$ – baseline	91.7	91.7	91.7
$\text{BERT}_{\text{large}}$ – W-Avg	91.9	92.0	92.0
$\text{BERT}_{\text{large}}$ – 2×W-Avg	92.5	92.5	92.5
$\text{BERT}_{\text{large}}$ – 2×W-Avg + MT	92.8	92.7	92.8

Table 4: Results in terms of micro-averaged precision, recall and F1 score on SRL over the verbal predicate instances in the standard gold benchmark of CoNLL-2009 for dependency-based SRL.

Results on SRL. Table 4 compares the results obtained on the verbal predicate instances in the standard gold benchmark of CoNLL-2009 for dependency-based SRL.[5] As we can see, each contribution provides an improvement over the previous one, both when using BERT-base-cased and BERT-large-cased (+0.4% and +1.1% in F1 score[6] over the baseline, respectively), the latter being one of the most used pretrained language models to achieve state-of-the-art results on the task. In general, not only did our analysis shed light on interesting properties of current PLMs through the lens of predicate senses and their rolesets, but it also provided practical hints on how to better exploit such properties in SRL.

Qualitative Analysis. Finally, we provide a look at what happens when our model is informed about predicate senses and their rolesets at training time. To inspect how the vector representations of predicates change as we inject more inductive bias towards predicate-argument information, in Figure 3 we use t-SNE to project and visualize on a bidimensional plane the representations of the predicate *close* when using: i) the baseline model, which is unaware of predicate-argument information and, therefore, does not show any significant clustering according to different rolesets; ii) the model when it can use different weighted averages to com-

[5] We trained our model for 30 epochs using Adam with an initial learning rate of 1e-3, leaving all parameters of the underlying language model frozen and using the parameter values used in the original paper by Conia and Navigli (2020).

[6] Scores were computed using the official CoNLL-2009 scorer provided during the shared task. This scoring script produces a unified F1 measure that takes into account both predicate senses and semantic roles.

Figure 3: t-SNE visualization of the representations for the predicate *close*. Different colors represent different rolesets, even though some rolesets are partially overlapping (e.g. {AM-EXT, AM-MNR} and {AM-EXT, AM-TMP}). From left to right: predicate representations from the baseline SRL model which is completely unaware of rolesets (left); predicate representations from an SRL model that can use two different weighted averages to create different representations for predicate senses and their arguments (center); predicate representations from an SRL model that is tasked to explicitly identify rolesets through a secondary learning objective in a multi-task fashion (right).

pute representations for predicate senses and their arguments; and iii) the model when it is explicitly tasked with the secondary training objective of learning to identify the roleset of each predicate. As one can see, as we inject more linguistic information into the model, the representations can be clustered better according to their corresponding predicate-argument structures.

5 Conclusion

In this paper, we probed PLMs for PASs: differently from past work, we dissected SRL into its core subtasks and analysed how PLMs encode predicate-argument structure information such as predicate senses and their rolesets. In our analysis, we observed that, despite the intrinsic connection between predicate senses and their rolesets that exists in several popular SRL inventories, different PLMs encode their features across significantly different layers. What is more, we also discovered that verbal and nominal predicates and their PASs are represented differently, making verbal-to-nominal SRL transfer far from trivial, and providing an empirical explanation for why previous attempts in this direction have struggled to obtain strong results. Furthermore, our analysis revealed that current multilingual language models encode PASs similarly across two very different languages, namely, English and Chinese.

Finally, in contrast to previous work on probing, we put together what we learned and demonstrated a practical application of our findings by devising simple yet effective techniques for the integration of predicate-argument structure knowledge into a state-of-the-art end-to-end architecture for SRL.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Consolidator Grant MOUSSE No. 726487 and the European Language Grid project No. 825627 (Universal Semantic Annotator, USeA) under the European Union's Horizon 2020 research and innovation programme.

This work was supported in part by the MIUR under grant "Dipartimenti di Eccellenza 2018-2022" of the Department of Computer Science of Sapienza University of Rome.

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, COLING-ACL '98, August 10-14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference*, pages 86–90. Morgan Kaufmann Publishers / ACL.

Rexhina Blloshmi, Simone Conia, Rocco Tripodi, and Roberto Navigli. 2021. Generating senses and roles: An end-to-end model for dependency- and span-based Semantic Role Labeling. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3786–3793. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2753–2765, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rui Cai and Mirella Lapata. 2019. Syntax-aware Semantic Role Labeling without parsing. *Transactions of the Association for Computational Linguistics (TACL)*, 7:343–356.

Rui Cai and Mirella Lapata. 2020. Alignment-free cross-lingual Semantic Role Labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3883–3894, Online. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. Unifying cross-lingual Semantic Role Labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351, Online. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual Semantic Role Labeling: A language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

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

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: A novel large-scale verbal semantic resource and its application to Semantic Role Labeling. 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 627–637, Hong Kong, China. 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.

Shexia He, Zuchao Li, and Hai Zhao. 2019. Syntax-aware multilingual Semantic Role Labeling. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5349–5358. 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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or span, end-to-end uniform Semantic Role Labeling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6730–6737. AAAI Press.

Diego Marcheggiani and Ivan Titov. 2020. Graph convolutions over constituent trees for syntax-aware Semantic Role Labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3915–3928, Online. Association for Computational Linguistics.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Special issue introduction: Semantic Role Labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Roberto Navigli. 2018. Natural Language Understanding: Instructions for (present and future) use. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5697–5702. ijcai.org.

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

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. 2018. Searching for activation functions. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*.

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

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. 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.

Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. 2020. A cross-task analysis of text span representations. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 166–176, Online. Association for Computational Linguistics.

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

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yanpeng Zhao and Ivan Titov. 2020. Unsupervised transfer of semantic role models from verbal to nominal domain. *CoRR*, abs/2005.00278.