Analysing Syntactic and Semantic Features in Pre-trained Language Models in a Fully Unsupervised Setting

Necva Bölücü
Computer Engineering
Hacettepe University
Adana Alparslan Turkes
Science and Technology University
Adana, Turkey
nbolucu@atu.edu.tr

Burcu Can
RGCL, University of Wolverhampton
Wolverhampton, UK
b.can@wlv.ac.uk

Abstract

Transformer-based pre-trained language models (PLMs) have been used in all NLP tasks and resulted in a great success. This has led to the question of whether we can transfer this knowledge to syntactic or semantic parsing in a completely unsupervised setting. In this study, we leverage PLMs as a source of external knowledge to perform a fully unsupervised parser model for semantic, constituency and dependency parsing. We analyse the results for English, German, French, and Turkish to understand the impact of the PLMs on different languages for syntactic and semantic parsing. We visualize the attention layers and heads in PLMs for parsing to understand the information that can be learned throughout the layers and the attention heads in the PLMs both for different levels of parsing tasks. The results obtained from dependency, constituency, and semantic parsing are similar to each other, and the middle layers and the ones closer to the final layers have more syntactic and semantic information.

1 Introduction

Transformer-based pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2019) have shown state-of-art performance in many down-stream NLP tasks. The performance of such large PLMs has also begged the question of what type of information that these models can naturally acquire through self-supervised learning. This has been investigated especially through probing tasks to analyse the linguistic information that is learned during pre-training of such large models (Liu et al., 2019a; Clark et al., 2019; Kovaleva et al., 2019; Pimentel et al., 2020; Rogers et al., 2020). One type of linguistic information that has been affluently analysed is syntactic information, and most of the recent probing studies have been based on this question: "Can transformer-based large language models learn syntactic structures during pre-training?". Recent studies address this question and propose unsupervised models that use syntactic knowledge obtained from PLMs for NLP tasks such as constituency (Kim et al., 2020a,b; Zeng and Xiong, 2022) and dependency parsing (de Lhoneux et al., 2022).

There are two aims in this study: 1. We aim to analyse the linguistic information that is learned by PLMs in different syntactic levels (dependency, constituency and semantic parsing) which deviates from the previous work, and provide a comparison with different languages. 2. We aim to demonstrate whether it is possible to use the linguistic information learned from PLMs in a fully unsupervised model for dependency, constituency and semantic parsing.

Existing approaches that use pre-trained language models are evaluated mainly on constituency parsing (Kim et al., 2020a,b) and dependency parsing (Hewitt and Manning, 2019; Clark et al., 2019). However, there is not any study that evaluates various parsing levels including semantic parsing using the same parsing model and compares the parsing results to understand the behaviour of PLMs for different levels of parsing from syntax to semantics.

In this paper, we evaluate a fully unsupervised model for three parsing tasks. We adopt the chart-based zero-shot parsing model (Kim et al., 2020b) that is based on the syntactic distance concept (Shen et al., 2017, 2018). To our knowledge, this will be the first study that combines syntactic distance with PLMs to apply to semantic parsing with zero-shot learning. In this study, we particularly use UCCA graph-based semantic representation for semantic parsing, which has been tackled as a constituency parsing problem in previous studies (Jiang et al., 2019; Bölücü and Can, 2021). In addition to the well-studied languages such as English, German, and French, we also evaluate

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 19 - 31
December 15-18, 2022

the models for Turkish with a comparably smaller dataset. We obtain the best results with multilingual PLMs. The results show that the zero-shot parsing model performs better with shorter sentences. It also shows that PLMs performs the best with middle layers and the ones closer to the final layers interestingly for all of the three parsing tasks, which are in line with the previous studies.

2 Related Work

A recent research direction has been towards analysing PLMs without fine-tuning on a particular down-stream task to explore the type of information that is learned during pre-training, which is called *probing*. For that purpose, PLMs have been investigated in various tasks such as language modeling (Shen et al., 2017), dependency parsing (Hewitt and Manning, 2019; Clark et al., 2019), constituency parsing (Shen et al., 2017; Peters et al., 2018; Li et al., 2020; Kim et al., 2020b), discourse parsing (Wu et al., 2020), commonsense reasoning (Tikhonov and Ryabinin, 2021), and grammar induction (Shen et al., 2018; Kim et al., 2020a). Some of the studies have also questioned if the syntax is encoded in PLMs (Shi et al., 2016; Blevins et al., 2018; Jawahar et al., 2019) and some of them analysed how large language models encode other types of linguistic information such as coference, entity information, parsing, and semantic roles, and NER (Tenney et al., 2019; Liu et al., 2019a).

In line with our work, Shen et al. (2017) use syntactic distance for character-level and word-level language modeling, and unsupervised constituency parsing. Li et al. (2020) use PLMs for unsupervised constituency parsing focusing on attention heads by ranking and ensembling them. Kim et al. (2020b) propose a model with chart-based decoder for the same problem, which also solves the greedy search problem of Shen et al. (2018). All of these studies are based on the idea that the syntactic structure of sentences are naturally learned along with language modeling. Some of those works (Kim et al., 2020a; Wu et al., 2020) also combine syntactic distance with PLMs to induce syntactic structure in an unsupervised setting. A recent work by Shen et al. (2021a) also introduces joint learning of constituency parsing with dependency parsing in an unsupervised framework.

Our work is similar to these and we also follow the chart-based zero-shot parsing introduced by Kim et al. (2020b). However, this is the first time that several parsing tasks are tackled using the same unsupervised model and this is the first time UCCA-based semantic parsing is performed in an unsupervised setting.

3 Chart-based Zero-shot Parsing

We utilise the syntactic distance concept (Shen et al., 2017, 2018) which was particularly explored for constituency parsing (Kim et al., 2020a,b; Li et al., 2020) by directly using the PLMs without fine-tuning. Here, we adopt chart-based zero-shot parsing based on syntactic distance for three different parsing problems that are semantic, constituency and dependency parsing to explore usability of PLMs in zero-shot setting.

The method calculates scores for spans where an input sentence $s = \{w_1, \cdots, w_n\}$ is made up of a set of labeled spans as follows:

$$T = \{(i_t, j_t, l_t) : t = 1, \cdots, |T|\}$$

where i_t and j_t refer to the beginning and ending positions of the t^{th} span respectively with the label set $l_t \in L$. A score $s(t)$ is assigned to each tree, which is decomposed as follows:

$$s(t) = \sum_{(i,j) \in t} s_{span}(i, j) \quad (1)$$

Here, $s_{span}(i, j)$ denotes per-span scores that are calculated recursively by splitting spans into smaller spans as defined below:

$$s_{split}(i, k, j) = s_{span}(i, k) + s_{span}(k + 1, j)$$
$$s_{span}(i, j) = s_{comp}(i, j) + min_{i \leq k < j} s_{split}(i, k, j)$$

where $s_{comp}(\cdot, \cdot, \cdot)$ measures the validity of the compositionality of the $span(i, j)$ itself, while $s_{split}(i, k, j)$ indicates how plausible it is to split span (i, j) at position k. To calculate $s_{comp}(\cdot, \cdot, \cdot)$, Kim et al. (2020b) introduced two alternative labeled functions. The first one is the characteristic score function $s_c(\cdot, \cdot)$, and the second is the pair score function $s_p(\cdot, \cdot)$. Pair score function computes the average pairwise distance in a given span:

$$s_p(i, j) = \frac{1}{\binom{j-i+1}{2}} \sum_{(w_x, w_y) \in pair(i,j)} f(g(w_x), g(w_y))$$

$$s_c(i, j) = \frac{1}{j - i + 1} \sum_{i \leq x \leq j} f(g(w_x), c)$$

$$c = \frac{1}{j - i + 1} \sum_{i \leq y \leq j} g(w_y)$$

where $pair(i, j)$ returns a set of all combinations of bigrams (e.g. w_x, w_y) inside the span (i, j). Function $f(\cdot, \cdot)$ is the distance measure and $g(\cdot)$ is the representation function. Jensen-Shannon (JSD) and Hellinger (HEL) distance functions are used to measure the distance between two spans. $g = \{g^d_{(u,v)} | u = 1, \cdots, l, v = 1, \cdots, a\}$ returns the v^{th} attention head on the u^{th} layer of the pre-trained language model.

CYK (Cocke-Younger-Kasami) (Chappelier and Rajman, 1998) is used for decoding to generate the trees. The parser outputs tree $\hat{t}$ that has the lowest score:

$$\hat{t} = \arg\min_{T} s(t) \tag{2}$$

For each distance function with score functions , we obtain the weights of the i^{th} layer and j^{th} attention head of that layer. Then we calculate the span scores using the distance functions. We select the tree with the lowest score for each distance function, which leads to 4 trees in i^{th} layer and j^{th} attention head. Therefore, we finally obtain $4 \times l \times a$ trees, where l is the number of layers and a is the number of attention heads. The final F1 scores are calculated for each tree and the highest F1 score is reported in the results.

4 Three Levels of Parsing with a Single Model

We use the chart-based zero-shot parsing model for three types of parsing ranging in different semantic and syntactic levels with different granularities and structures of a given text:

Dependency Parsing Dependency parsing is concerned with the syntactic relations between words in a sentence. Those syntactic relations are discovered in terms of dependencies of words on each other. In order to apply the zero-shot parsing model for dependency parsing, we compute the scores for each tree and then we apply Eisner (Eisner, 1996) decoding algorithm (rather than CYK) to produce dependency trees using the tree scores.

Constituency Parsing Constituency parsing is concerned with extracting the syntactic structure of a given text through phrasal constituents. Therefore, it is more concerned with the syntactic structure of an entire sentence rather than the relations between words as opposed to dependency parsing. We apply zero-shot parsing without adding any additional step for constituency parsing.

Semantic Parsing Semantic parsing is concerned with extracting the semantic structure of a given text using a formal representation. We particularly use UCCA (Abend and Rappoport, 2013) graph-based semantic representation to extract semantic relations within the text. In order to perform UCCA-based parsing, we first convert UCCA graphs into constituent trees by removing discontinuities and remote edges (Jiang et al., 2019; Bölücü and Can, 2021). Then we perform zero-shot learning to tackle semantic parsing as a constituency parsing problem. After finding the tree with the lowest score, we convert constituency trees back into the UCCA-based graphs, and restore discontinuity units. We disregard the remote edges and implicit edges.

5 Experiments and Results

We conducted experiments to evaluate the unsupervised parser on dependency, constituency, and semantic parsing for English, German, French, and Turkish since UCCA-annotated datasets are only available for these languages.

5.1 Datasets

- **Dependency Parsing:** We used Universal Dependency v2.3 (Schuster et al., 2017) datasets in English, German, French and Turkish.
- **Constituency Parsing:** We used Penn Treebank (PTB) (Marcinkiewicz, 1994) for English, the SPMRL dataset (Seddah et al., 2013) for German and French, and the Turkish Annotated Treebank (Yıldız et al., 2016) for Turkish.
- **Semantic Parsing:** We used UCCA datasets provided by SemEval 2019 (Hershcovich et al., 2019) in English, German, and French, and the Turkish UCCA-annotated dataset released by Bölücü and Can (2022).

Since it is a zero-shot parsing model and does not involve a training stage, we only used the test sets[1] for all languages for the evaluation.

5.2 Experimental Setting

We use both monolingual and multilingual PLMs in the experiments. For English, we use the following monolingual PLMs: BERT (Devlin et al., 2019),

[1]The details of the datasets are given in Table 7 in Appendix A.

GPT-2 (Radford et al., 2019), RoBERTa (Liu et al., 2019b), and XLNet (Yang et al., 2019). We follow previous work (Kim et al., 2020a,b; Li et al., 2020) by using two variants of each PLM, where the X-base variant consists of 12 layers, 12 attention heads and 768 hidden dimensions, while the X-large variant has 24 layers, 16 attention heads and 1024 hidden dimensions. GPT2 model corresponds to X-base while GPT2-medium corresponds to X-large model.

We use `bert-base-german-cased`, `bert-base-french-europeana-cased`, and `bert-base-turkish-cased` for German, French and Turkish monolingual PLMs respectively.

For multilingual experiments, we use multilingual version of the BERT-base model (M-BERT) (Devlin et al., 2019), the XLM-base model (XLM-R[2]) (Conneau and Lample, 2019), which is a multilingual RoBERTa model, and the large version of XLM (XLM-R-large) (Conneau et al., 2020).

5.3 Results

We present the results obtained from each parsing separately below[4].

Dependency Parsing Dependency parsing results for all languages are given in Table 1. The best results are obtained from multilingual PLMs in all languages. Since the other unsupervised dependency parsing models are either finetuned (Ma and Xia, 2014; Shen et al., 2021b) or utilise other external resources such as Google Universal Treebanks (Ma and Xia, 2014) or WSJ (Shen et al., 2021b), we have not made a comparison with other models since the model presented here is fully unsupervised, does not use any annotated data, and does not incorporate any syntactic information during PLM pre-training.

Constituency Parsing For constituency parsing, we either perform top-down or chart-based parsing to generate trees. We further experiment with using different layers in the PLMs. All unlabeled F1 scores for the constituency parsing are given in Table 2 and Table 3. We use abbreviations TD, CP, and CC for Top-Down, Chart-Pair (pair score function $s_p(\cdot,\cdot)$) and Chart-Characteristic (characteristic score function $s_c(\cdot,\cdot)$) respectively. Except English, we obtain the best results with the top-down decoder and with XLM-R for German, French, and Turkish [5].

Semantic Parsing The unlabeled F_1 scores for UCCA-based semantic parsing are given in Table 4 and 5. The best results are obtained from RoBERTa-base amongst the monolingual models and from XLM-R amongst the multilingual models in English. Interestingly, both RoBERTa and XLM-R gives similar results. For German, French, and Turkish, all the best results are obtained from multilingual models. Since this is the very first study that performs UCCA-based semantic parsing in a completely unsupervised framework, there is not any other study that is available to compare with ours. Therefore, we report our results only as the baseline results for the future studies.

Dependency parsing scores are comparably much lower than both constituency and semantic parsing in all languages. Unsupervised dependency parsing has been mostly performed using probabilistic generative models in the literature (Klein and Manning, 2004) and it is comparatively harder than constituency parsing since it requires learning finer relations between words rather than phrases in a sentence. However, interestingly, UCCA-based semantic parsing scores are also promising and as good as constituency parsing performance. It should be noted that UCCA-based semantic parsing has not been tackled with an unsupervised learning model before.

As for the PLM models, the GPT and GPT2-medium perform comparatively poorly on all parsing problems. Unlike other PLMs, the GPT models are auto-regressive language models that do not allow to incorporate the context on both sides of a word, which might be the reason of the poor performance of the GPT models.

5.4 Analysis of the Results

We analyse the attention layers and heads that contribute the most to each parsing task, along with the affect of the sentence length in the experiments.

5.4.1 Attention Layers

We analyse the attention layers to see which layers provide the most information for the parsing tasks.

[2]The details of the training datasets used in the experiments are given in Table 8 in Appendix B.

[3]We used the pre-trained models of BERT defined in Section A for each language.

[4]We give the results of supervised models in Appendix C.

[5]The model is adopted from that of Kim et al. (2020b) and we prefer not to repeat the comparative scores here again.

Model	*Monolingual Models*			
	English	German	French	Turkish
BERT-base-cased[3]	26.48	26.59	24.78	35.56
BERT-large-cased	27.89	-	-	-
XLNet-base-cased	25.66	-	-	-
XLNet-large-cased	27.53	-	-	-
RoBERTa-base	27.68	-	-	-
RoBERTa-large	25.11	-	-	-
GPT2	19.66	-	-	-
GPT2-medium	21.44	-	-	-
PLM	*Multilingual Models*			
M-BERT	30.80	30.69	**34.37**	**41.62**
XLM-R	30.80	**31.84**	34.27	41.25
XLM-R-large	**32.66**	28.58	26.19	39.13

Table 1: UAS scores for dependency parsing.

	English			German		
Model	TD	CP	CC	TD	CP	CC
	Monolingual Models					
BERT-base-cased	34.51	40.24	42.05	26.96	24.82	26.59
BERT-large-cased	38.93	43.68	44.58	-	-	-
XLNet-base-cased	40.12	42.14	43.47	-	-	-
XLNet-large-cased	38.32	42.60	43.73	-	-	-
RoBERTa-base	40.61	45.37	**46.01**	-	-	-
RoBERTa-large	34.30	42.19	43.26	-	-	-
GPT2	34.21	34.01	35.78	-	-	-
GPT2-medium	37.65	38.59	39.81	-	-	-
	Multilingual Models					
M-BERT	40.28	43.44	44.13	30.69	30.59	30.28
XLM-R	41.25	44.25	44.76	**33.13**	32.19	31.84
XLM-R-large	39.13	42.87	44.67	28.18	27.13	28.58

Table 2: Unlabeled F1 scores for constituency parsing in English and German.

	French			Turkish		
Model	TD	CP	CC	TD	CP	CC
	Monolingual Models					
BERT-base-cased	24.78	22.83	23.86	35.36	31.47	33.50
	Multilingual Models					
M-BERT	32.88	30.37	30.45	41.29	40.61	39.93
XLM-R	**34.19**	31.29	30.93	**45.18**	43.49	42.30
XLM-R-large	26.68	25.70	26.46	36.21	36.72	36.72

Table 3: Unlabeled F1 scores for constituency parsing in French and Turkish.

Dependency Parsing UAS scores obtained from multilingual models for each layer are illustrated in Figure 1. The results show that we get the highest UAS scores from the middle or the ones closer to the final layers of PLMs for all languages.

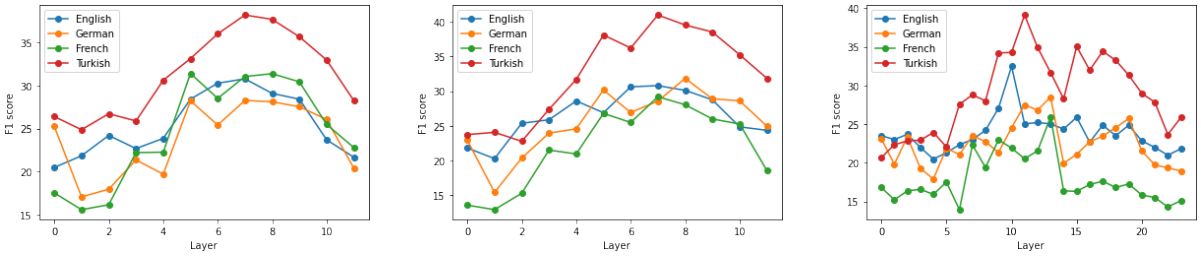
Constituency Parsing F1 scores obtained from multilingual PLMs for all layers are given in Figure 2. Although there are slight differences between languages, the general picture does not differ from the dependency parsing results and again the highest scores are obtained from mostly middle

	English-Wiki			English-20K		
Model	TD	CP	CC	TD	CP	CC
	Monolingual Models					
BERT-base-cased	38.34	42.30	42.60	39.10	42.80	43.93
BERT-large-cased	38.33	42.93	43.52	39.41	43.82	44.75
XLNet-base-cased	37.00	39.62	40.18	37.57	39.56	42.70
XLNet-large-cased	38.41	41.25	42.27	39.98	41.20	41.52
RoBERTa-base	41.82	44.96	**45.21**	32.43	45.62	**46.18**
RoBERTa-large	37.65	41.37	41.62	36.44	41.78	41.92
GPT2	31.97	38.23	38.56	32.41	37.97	38.40
GPT2-medium	34.86	38.49	38.58	32.21	37.68	39.31
	Multilingual Models					
M-BERT	39.62	43.52	44.06	38.11	43.99	45.15
XLM-R	40.98	45.45	**45.89**	42.06	45.51	**46.30**
XLM-R-large	36.40	40.05	40.87	33.69	40.00	41.45

Table 4: Unlabeled F1 scores for semantic parsing in English (English-Wiki, English-20K).

	German-20K			French-20K			Turkish		
Model	TD	CP	CC	TD	CP	CC	TD	CP	CC
	Monolingual Models								
BERT-base-cased[3]	40.30	41.93	42.96	40.32	40.55	42.71	41.49	39.50	42.15
	Multilingual Models								
M-BERT	39.08	**44.17**	44.07	41.01	43.26	46.08	42.15	44.80	44.14
XLM-R	40.90	43.15	42.98	44.13	46.08	**47.38**	46.79	**48.77**	46.79
XLM-R-large	35.59	39.63	42.37	37.56	39.17	38.94	45.46	44.14	46.13

Table 5: Unlabeled F1 scores for semantic parsing in German, French and Turkish.



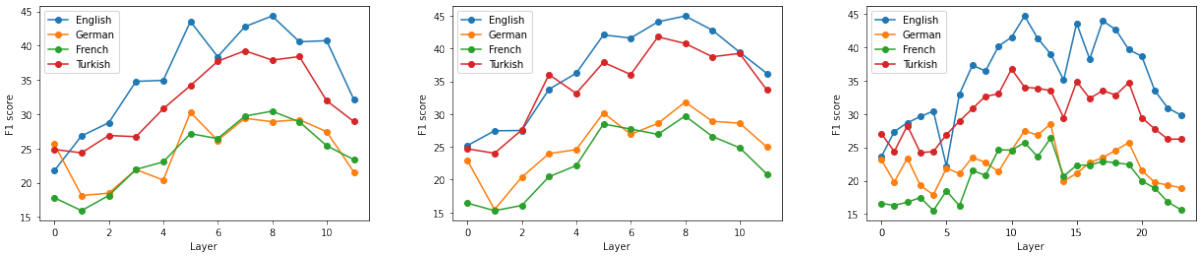
(a) Bert-multilingual (b) XLM-R (c) XLM-R-large

Figure 1: UAS scores of multilingual PLMs for dependency parsing.

layers.

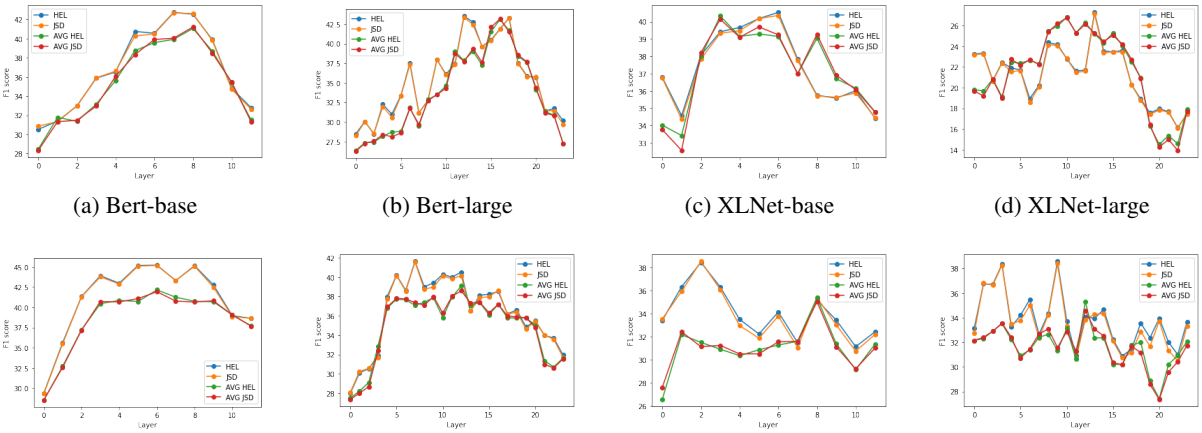
Semantic Parsing F1 scores obtained from monolingual models for all layers along with different distance functions are given in Figure 3. Only the best scores obtained from the attentions in each layer are illustrated. The graphs show that there is not much difference between the distance functions in terms of their performance in parsing. However, we obtain the highest scores again from the middle or towards the last layers except for GPT-2, which achieves the best in the lower layers.

The results obtained from multilingual PLMs for all languages are given in Figure 4. The F1 scores of languages are very low in the first hidden layers except for Turkish. The lower hidden layers might be more informative in short sentences because the Turkish UCCA dataset involves shorter sentences compared to other languages. This might be the reason of such a difference between the languages. The results also support that the final layers bear more syntactic information compared to the lower



(a) Bert-multilingual (b) XLM-R (c) XLM-R-large

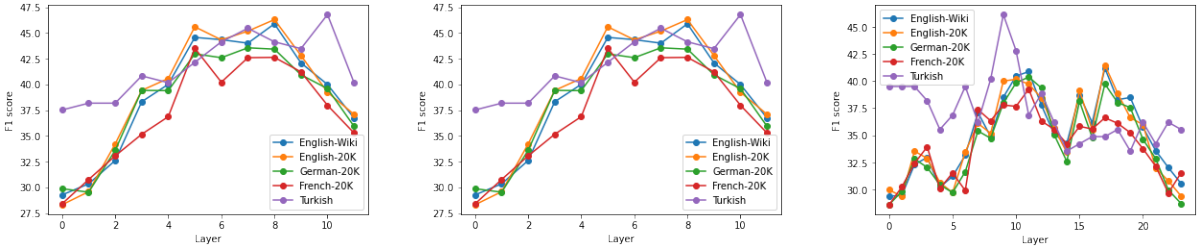
Figure 2: F1 scores of multilingual PLMs for constituency parsing.



(a) Bert-base (b) Bert-large (c) XLNet-base (d) XLNet-large

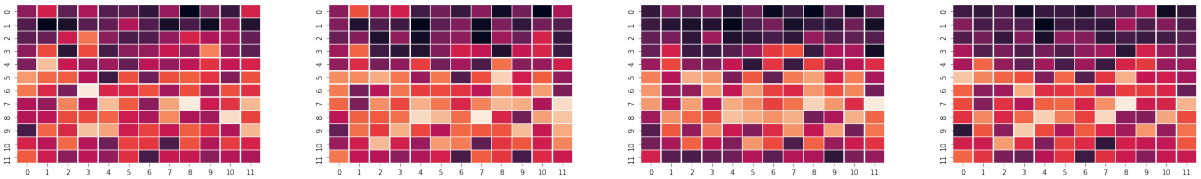
(e) RoBERTa-base (f) RoBERTa-large (g) GPT2 (h) GPT2-medium

Figure 3: F1 scores from monolingual PLMs using the English Wiki dataset for semantic parsing.



(a) Bert-multilingual (b) XLM-R (c) XLM-R-large

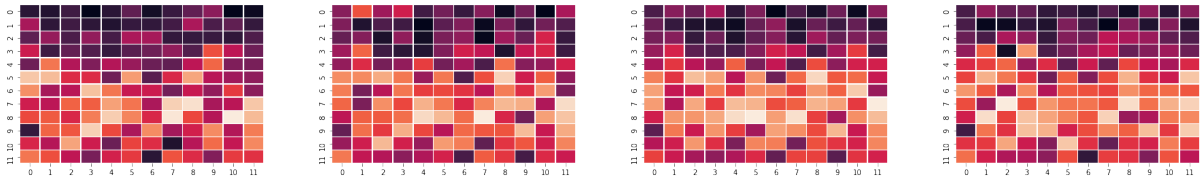
Figure 4: F1 scores from multilingual PLMs using the UCCA datasets.



(a) English (b) German (c) French (d) Turkish

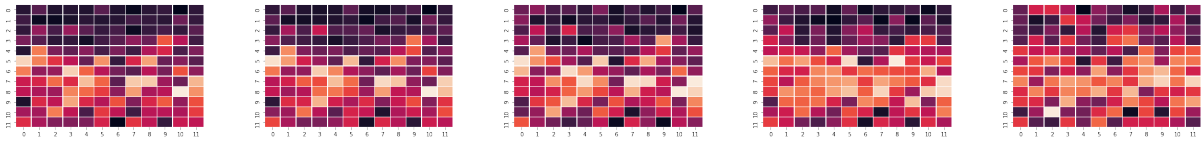
Figure 5: Unsupervised dependency parsing performance in all languages according to different attention heads and hidden layers with HEL distance function (Light cells refer to higher UAS scores).

layers, especially in longer sentences, which is consistent with the findings of other studies (Clark et al., 2019; Kim et al., 2020b,a).



(a) English (b) German (c) French (d) Turkish

Figure 6: Unsupervised constituency parsing performance in all languages for different attention heads and hidden layers with HEL distance function (Light cells correspond to higher F1 scores).



(a) English-Wiki (b) English-20K (c) German-20K (d) French-20K (e) Turkish

Figure 7: Unsupervised UCCA semantic parsing performance in all languages with different attention heads and hidden layers with HEL distance function (Light cells refer to higher F1 scores).

5.4.2 Attention Heads

We also analyse the attention heads in the layers to observe which attention heads contribute the most to each parsing task. F1 scores obtained from the attention heads in different layers are given in Figure 5, Figure 6, and Figure 7 for dependency, constituency, and semantic parsing (with XLM-R) respectively. The graphs support the findings regarding the hidden layers and further show that top heads contain more information in all tasks and languages apart from Turkish constituency and semantic parsing where the lower heads contain more information. This might be again due to the length of the sentences in the Turkish datasets.

5.4.3 Sentence Length

To understand the effect of the sentence length, we extract the average length of the sentences in all datasets. The average sentence length of Turkish datasets for all tasks is less than that of the other languages, whereas the average sentence length of German and French is higher in all parsing datasets.

To investigate the relationship between the sentence length and the accuracy of the parsing, we run the constituency parsing with XLM-R multilingual PLM and top-down parser on 1000 samples with a length less than the average length of the dataset and 1000 samples with a length greater than the average length of the datasets in English, French and German. We only use 50 samples (25 less and 25 are greater than the average length in Turkish since there are only 63 samples in the dataset. Table 6 gives the average length of the sentences in each dataset along with the obtained F1 scores. The results show that the model performs better on shorter sentences. This also confirms that the model can hardly find distant relationships in longer sentences.

6 Conclusion

We analyse the syntactic information learned by transformer-based PLMs for various parsing problems (namely dependency, constituency, and semantic parsing) using a fully unsupervised zero-shot parser. To the best of our knowledge, this is the first study that compares an unsupervised model for three different parsing problems in a fully unsupervised setting and analyses the linguistic information learned from PLMs during pre-training for three different parsing tasks from syntax to semantics. The results show that PLMs provide information from mostly middle and towards the final layers for all parsing tasks, which is also in line with the previous work on constituency and dependency parsing. However, interestingly, the study shows that when it comes to structure learning, syntax and semantics are both encoded in middle and towards the final layers.

References

Omri Abend and Ari Rappoport. 2013. UCCA: A semantics-based grammatical annotation scheme. In *Proceedings of the 10th Interna-*

		samples< average		samples>average	
Language	Av. Len	Av. Len	F1	Av. Len	F1
English	20.46	12.76	43.73	28.96	39.04
German	18.82	11.62	33.71	30.07	30.37
French	29.73	16.82	34.64	46.05	31.85
Turkish	13.95	11.84	48.42	16.48	44.29

Table 6: F1 scores of constituency parsing for samples that are shorter and longer than the average overall sentence length.

tional Conference on Computational Semantics (IWCS 2013)–Long Papers, pages 1–12.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. *CoRR*, abs/1805.04218.

Necva Bölücü and Burcu Can. 2022. Turkish universal conceptual cognitive annotation. In *Proceedings of LREC 2022*. European Language Resources Association.

Necva Bölücü and Burcu Can. 2021. Self-attentive constituency parsing for UCCA-based semantic parsing. *arXiv preprint arXiv: 2110.00621*.

J-C Chappelier and Martin Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG. In *Proceedings of the 1st Workshop on Tabulation in Parsing and Deduction (TAPD'98)*, CONF, pages 133–137.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. *arXiv preprint arXiv:2203.08555*.

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.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for UCCA. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1138.

Daniel Hershcovich, Zohar Aizenbud, Leshem Choshen, Elior Sulem, Ari Rappoport, and Omri Abend. 2019. SemEval-2019 Task 1: Cross-lingual semantic parsing with UCCA. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1–10.

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.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Wei Jiang, Zhenghua Li, Yu Zhang, and Min Zhang. 2019. HLT@ SUDA at SemEval-2019 Task 1: UCCA graph parsing as constituent tree parsing. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 11–15.

Taeuk Kim, Jihun Choi, Sang-goo Lee, and Daniel Edmiston. 2020a. Are Pre-trained Language Models Aware of Phrases? Simple but Strong Baselines for Grammar Induction. In *International Conference of Learning Representations*.

Taeuk Kim, Bowen Li, and Sang-goo Lee. 2020b. Multilingual chart-based constituency parse extraction from pre-trained language models. *arXiv preprint arXiv:2004.13805*.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 478–485, Barcelona, Spain.

Daniel Kondratyuk. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *CoRR*, abs/1904.02099.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. *arXiv preprint arXiv:1908.08593*.

Bowen Li, Taeuk Kim, Reinald Kim Amplayo, and Frank Keller. 2020. Heads-up! unsupervised constituency parsing via self-attention heads. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 409–424, Suzhou, China.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *CoRR*, abs/1903.08855.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1337–1348.

Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Using Large Corpora*, page 273.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

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*.

Sebastian Schuster, Matthew Lamm, and Christopher D. Manning. 2017. Gapping constructions in universal dependencies v2. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 123–132, Gothenburg, Sweden.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, et al. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182. Association for Computational Linguistics.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2017. Neural language modeling by jointly learning syntax and lexicon. *arXiv preprint arXiv:1711.02013*.

Yikang Shen, Zhouhan Lin, Athul Paul Jacob, Alessandro Sordoni, Aaron Courville, and Yoshua Bengio. 2018. Straight to the tree: Constituency parsing with neural syntactic distance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1180, Melbourne, Australia. Association for Computational Linguistics.

Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Don Metzler, and Aaron Courville. 2021a. Structformer: Joint unsupervised induction of dependency and constituency structure from masked language modeling. In *ACL 2021*.

Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, and Aaron Courville. 2021b. StructFormer: Joint unsupervised induction of dependency and constituency structure from masked language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7196–7209, Online.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1526–1534.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *CoRR*, abs/1905.05950.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *arXiv preprint arXiv:2106.12066*.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Olcay Taner Yıldız, Ercan Solak, Şemsinur Çandır, Razieh Ehsani, and Onur Görgün. 2016. Constructing a Turkish constituency parse treebank. In *Information Sciences and Systems 2015*, pages 339–347. Springer.

Zhiyuan Zeng and Deyi Xiong. 2022. Unsupervised and few-shot parsing from pretrained language models. *Artificial Intelligence*, page 103665.

A Details of the test sets

Here specify the size of the test sets used in all parsing tasks.

	English	German	French	Turkish
DP	2077	1000	416	979
CP	2416	5000	2541	63
SP	Wiki: 515 20-K: 492	652	239	50

Table 7: Size of the test sets used in the experiments (DP: Dependency parsing, CP: Constituency parsing, and SP: Semantic parsing)

B Details of the training sets for XLM-R

Here we present the size of the monolingual datasets used for training the XLM-R.

Language	Tokens (M)	Size (GiB)
English	55608	300.8
German	10297	66.6
French	9780	56.8
Turkish	2736	20.9

Table 8: Size of each monolingual dataset used for training the XLM-R.

C Supervised model results for three levels of parsing

Here we give the results obtained from supervised models for dependency and semantic parsing problems with the best results of the unsupervised model in the paper[6].

Model	English	German	French	Turkish
Our Model	32.66	31.84	34.37	41.62
UDPipe ♣	89.63	85.53	90.65	74.19
UDify ♣	90.96	87.81	93.60	74.56

Table 9: Comparative UAF scores of our unsupervised model with supervised models for dependency parsing (♣: Kondratyuk (2019))

[6]We couldn't give the constituency parsing results since the studies on constituency parsing present only labeled scores.

[7]We used the zero-shot experimental results in the paper of (Bölücü and Can, 2022) for Turkish dataset.

Model	English-Wiki	English-20K	German-20K	French-20K	Turkish
Our Model	45.89	46.30	44.17	47.38	48.77
Tupa ♣	85.00	82.20	90.30	74.00	-
HLT@SUDA ♡	87.20	85.20	92.80	86.00	-
Self-Attentive ♠	89.60	87.69	94.10	86.00	76.80[8]

Table 10: Comparative unlabeled F-1 scores of our unsupervised model with supervised models for semantic parsing (♣: Hershcovich et al. (2017), ♡: Jiang et al. (2019), ♠: Bölücü and Can (2021))