Investigating the Role of Different Transformer Layers in Persian Language Models

Hirbod Gholamniaetakhsami hirgh815@student.liu.se

Supervisor: Jenny Kunz jenny.kunz@liu.se

1. Abstract

Large language Models(LLMs) have demonstrated remarkable problem-solving capabilities. Such models can provide countless possibilities to help with tackling real-world problems, hence such success led to various research contributions in this domain. However the rapid progress of LLMs raises some concerns about Quality of generated responses and ethical considerations. Therefore to effectively capitalize on such model capacities as well as ensure the relevance of response, there is a need to conduct a comprehensive evaluation of LLMs. This project aims to explore the linguistic knowledge and structure learned by Language models during pre-trainnig.This is done by training a classifier on a diagnostic tasks using the representations from different layers as input and investigating the performance of model by applying a group of metrics that measure the amount of information from different perspectives.

2. Introduction

The Large language models have shown significant capabilities across a broad spectrum of tasks, attracting attention and deployed in numerous downstream applications[1]. The works involving LLMs includes diverse topics such as diverse topics such as architectural innovations, better training strategies, context length improvements, fine-tuning, multi-modal LLMs, robotics, datasets, benchmarking, efficiency, and much more[2].

In particular study [1], provided a panoramic perspective on evaluation of Large Language Models. In this study, evaluation of LLMs is categorized into three major groups: knowledge and capacity evaluation, alignment evaluation, and safety evaluation. Knowledge and capacity evaluation focuses on assessing the models' ability to generate text similar to human. Alignment evaluation, aims to investigate the goodness of generated text in terms of aligning with human values and expectations. Finally, safety evaluation examines the potential risks associated with the use of LLMs, such as generation of misleading contents.

One way to approach knowledge and capability evaluation is by performing probing studies. Such studies usually aim to test model's understanding of certain concepts or its ability to generate relevant responses. This is done by designing specific tasks, which will help gaining insights into limitations of Language models, and contributing to improve these models. Chen and Ding investigated the ability of creative thinking through utilizing divergent association task[3]. Their method involve asking model to generate unrelated words and computing the semantic distance between them.

The distribution of linguistics information is the primary target for this project. A suitable direction of studies is investigation different layers of corresponding models in relevant domains. In [4], authors investigated four syntax tasks in the context of Deep Recurrent Neural Networks (RNN). They concluded a soft syntactic hierarchy emerges from the internal representations of the models. Similarly in studies[5,6], authors observe that different linguistic tasks are better predicted from different layers of neural models indicating a hierarchy exist from low-level to high-level tasks. In [7], authors experimented with seven languages through their corresponding framework. By applying global metrics and local metrics in their analysis, they found that while a hierarchical ordering exist in former, the latter does not guarantee the same. Also the results for models are very language dependent implying that the information distribution across layers may vary depending on the language and the data that the models are trained on and in other words, it shows a lack of generalizability.

In this project we investigate the emergence of syntactic information from layers following the framework developed in [7]. The framework is applied to two pair of tasks: 1- Part-of-speech (POS) tagging and 2- Syntactic ancestors prediction.

In the third section, the methodology for the project including the specifics of probing tasks and the measurement used is presented. In the fourth section, data and the language models under study are described and in section five, we present the results of our project. We present our conclusion in section six. Finally, section seven contains some potential extensions for this project.

3. Methodology

In this project, we aim to perform the following two pairs of probing tasks to investigate the correspondence between language properties and the network depth:

- Part-of-speech (POS) tagging: This task involves assigning each word in a text a grammatical category, such as noun, verb, adjective, etc. The two types of POS tags are:
    - **Most frequent tags (MFTs)**: These are the tags that are the most common for a given word form, regardless of the context. For example, the word "book" is most often a noun, so its MFT is noun.
    - **Non-most frequent tags (non-MFTs)**: These are the tags that are less common for a given word form, and depend on the context. For example, the word "book" can also be a verb, as in "book a flight", so its non-MFT is verb.
- Syntactic ancestors' prediction: This task involves predicting the relative position of a word's syntactic head (the word that governs its grammatical function) and its head's head (the grandparent) in the dependency tree. The two types of syntactic ancestors are:
    - **Parents (P)**: These words directly depend on another word in the sentence. For example, in the sentence "She likes the book", the word "likes" is the parent of "She" and "book".
    - **Grandparents (GP)**: These words depend on the head of another word in the sentence. For example, in the sentence "She likes the book that I gave her", the word "likes" is the grandparent of "that" and "her".

Metrics

The following measurements are used to describe the information distribution across layers of models

Global Baselined Probing (GBP): This metric measures the difference between the probe accuracy on a given layer and the baseline layer (the uncontextualized embedding layer).

$$GBP_i = Acc(l_i) - Acc(l_0) \quad (1)$$

Local Baselined Probing (LBP): This metric measures the difference between the probe accuracy on a given layer and the previous layer. This metric estimates the information gain when taking a step from one layer to the next.

$$LBP_i = Acc(l_i) - Acc(l_{i-1}) \quad (3)$$

Emergent Information (EMI): This metric measures the relative information gain of a given layer by dividing the LCP by the sum of LCP over all layers. This metric shows the layer's contribution to the overall emergent information within the model.

$$LCP_i' = \max(0, LCP_i) \quad (5)$$

$$EMI_i = \frac{LCP_i'}{\sum_{k=1}^{L} LCP_k'} \quad (6)$$

Global Conditional Probing (GCP): This metric measures the difference between the probe accuracy on the concatenation of a given layer and the baseline layer, and the baseline layer alone. This metric aims to capture the information that a layer contributes beyond the baseline.

$$GCP_i = Acc([l_i; l_0]) - Acc(l_0) \quad (2)$$

Local Conditional Probing (LCP): This metric measures the difference between the probe accuracy on the concatenation of a given layer and the previous layer, and the previous layer alone. This metric accounts for the exclusive information of the previous layer that is absent in the current layer.

$$LCP_i = Acc([l_i; l_{i-1}]) - Acc(l_{i-1}) \quad (4)$$

EMI, Baselined Control (EMI-BL): This metric is a simplified version of EMI that uses LBP instead of LCP. This metric does not control for information that was already present in the previous layer.

$$LBP_i' = \max(0, LBP_i) \quad (6)$$

$$\text{EMI} - \text{BL}_i = \frac{LBP_i'}{\sum_{k=1}^{L} LBP_k'} \quad (7)$$

4. Model and data

Models

In this experiment we are investigating the two probing tasks on English and Persian Bert models[8,9] in comparison. Both models are accessed via Transformer library[10].

The classifier used is feed-forward neural network with 64 hidden Units and over 10 epochs. The model was implemented in Google Tensorflow[11].

Data

For this experiment, the training data for this task was extracted from English and Persian Universal Dependencies treebank. In order to perform a better comparison, we first identified the most similar dataset in two steps:

1- Checking whether the datasets has similar contents, in terms of the category of text (i.e. movies, news …).
2- By encoding the sentences using multilingual Bert model[12], and computing the cosine similarity between each pair of sentences and computing the relevant statistics[1].

To compare the results from metrics, the two following strategies were used:

- **Max Layer**: This strategy identifies the layer that maximizes a certain metric for each task. If this layer is deeper for one task than the other is, the former task is considered higher in the hierarchy. This strategy focuses on the layer where the task-relevant information is most accessible.

- **Early Contributions**: This strategy looks at the contribution of the early layers (1, 1+2, 1+2+3) to the overall information gain. If this contribution is higher for one task than the other is, the former task is considered lower in the hierarchy. This strategy focuses on where the task-relevant information first emerges in the model.

5. Results

The results for Language models are presented in two sections:

Maximum layer

In table 1 and table 2 the maximum layers for the metrics for both probing tasks are visible. For part-of-speech tagging We can see the layer that maximizes the GBP for nonMFTs are deeper than MFTs for both languages as expected, Due to their frequent occurrence MFTs might be encoded earlier. For GCP nonMFT values are less than MFTs. In case of local metrics maximal values are less interpretive and this is likely because of baselining on previous layer.

If we look at the figure 1 and 2 containing distribution of global vs local for both languages, we can note that the overall shape for distribution of MFTs for both global score GBP and local score follow the same pattern. There are some occasional difference at middle layers. On the other hand for nonMFTs, it is evident that English model follow a smoother pattern.

	GBP		*GCP*		*LBP*		*LCP*		*EMI*		*EMI-BL*	
	MFT	noMFT	MFT	noMFT	MFT	noMFT	MFT	noMFT	MFT	noMFT	MFT	noMFT
en	4	7	10	8	2	1	1	1	1	1	2	1
fa	3	4	7	5	1	1	1	1	1	1	1	1

Table 1: Layer of maximum score for POS tagging task. Higher values in the hierarchy are highlighted.

	GBP		*GCP*		*LBP*		*LCP*		*EMI*		*EMI-BL*	
	P	GP	P	GP	P	GP	P	GP	P	GP	P	GP
en	4	6	5	7	2	1	1	1	1	1	1	1
fa	1	3	2	1	2	1	2	1	2	1	2	1

Table 1: Layer of maximum score for Syntactic ancestors' prediction task. Higher values in the hierarchy are highlighted.

[1] The results are available in "Similarity_test.ipnyb"

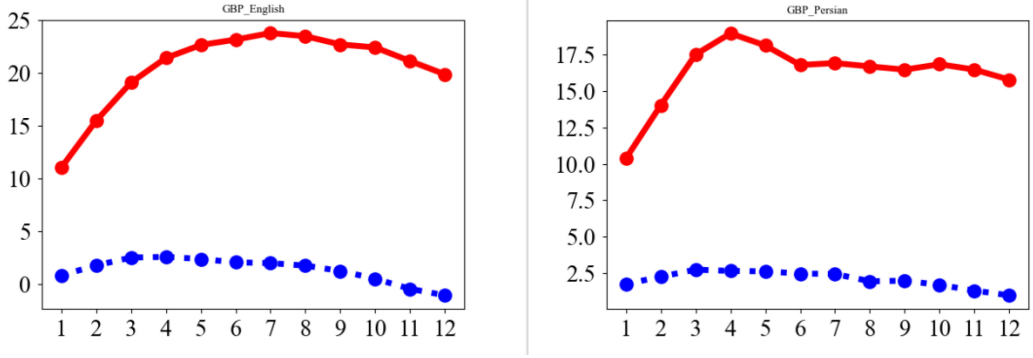

Figure 1: GBP scores for English vs Persian – blue for “MFT” ; red for “nonMFT”

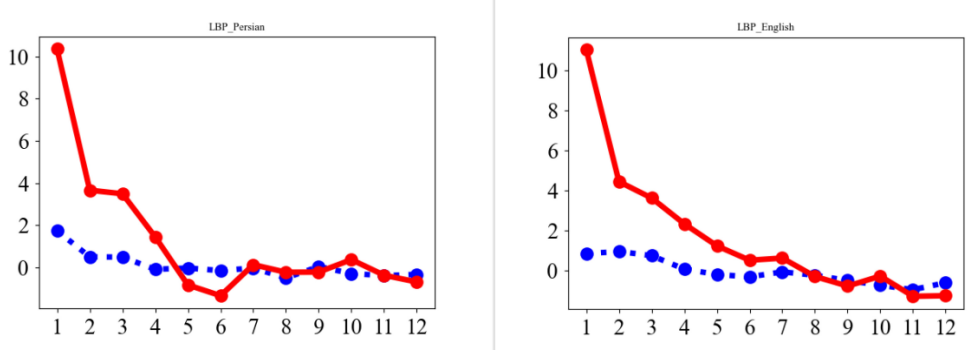

Figure 2: LBP scores for English vs Persian – blue for “MFT” ; red for “nonMFT”

In case of syntactic ancestors' prediction we can see that maximum layer GBP score belongs to Grandparents. Unlike English we can see that maximum GCP score belongs to second layer of parents. Interestingly, if we look at the figures 3 and figure 4 we can see that these languages follow different structure.

We note that the most notable differences in the trend of the following plots belong to the word’s parent, in the case of the GCP score for English we can see an increase in score between layers (2,4) whereas in the case of Persian, we note a gradual decrease in the score, while in the latter half of the plot we notice very few changes in the score. Looking at the local scores we may note that the trend and overall structure are very different, although the LBP score for English follows a smoother pattern in the word's parent.

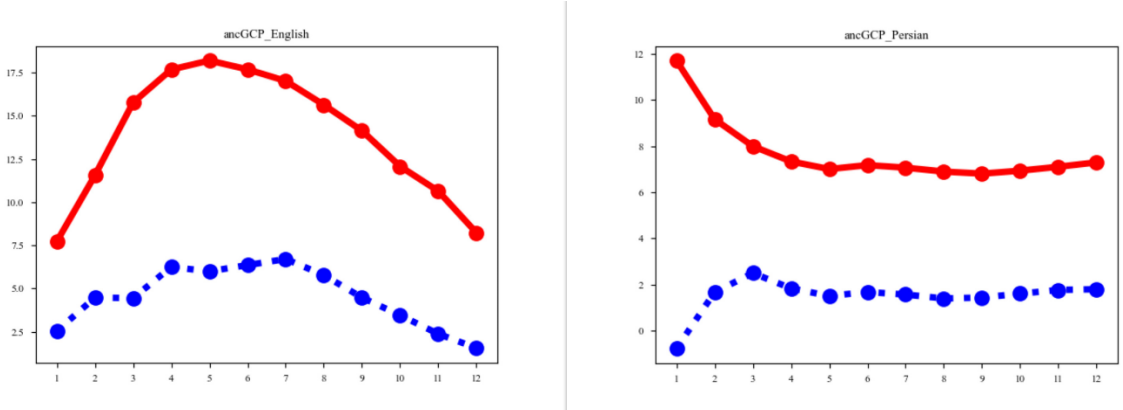

Figure 3: GCP scores for English vs Persian – blue for “GP” ; red for “P”

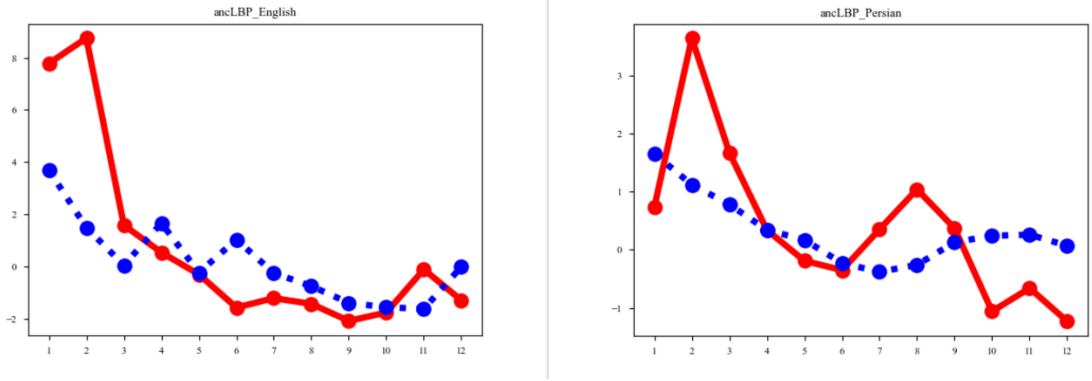

Figure 4: LBP scores for English vs Persian – blue for “GP” ; red for “P”

Early contributions

In table 3 and table 4, we can see the contribution of model in the first three layers for both of these probe tasks. In this case, we mostly rely on the EMI metric. If the contribution of earlier levels to the overall performance of the model is small, this could mean that the model does not rely much on these level to perform probing task.

In POS tagging task, In case of English we can see that MFT has smaller contribution to the first two layers than nonMFTs. If we look at the figure 5 we can observe the difference between early emergence of these two models. In this case, nonMFT has a smaller contribution in first three model for Persian, interestingly, the observable pattern for nonMFTs are more similar to nonMFTs in English model.

Regarding Syntactic Ancestors’ prediction task, for Persian model we can observe that word’s parent has bigger contribution than grandparents in early layers, which is the same pattern for English model, except for the second layer.

	Layer 1		*Layer 1+2*		*Layer 1+2+3*	
	MFT	noMFT	MFT	noMFT	MFT	noMFT
en	31.00	46.24	59.68	62.57	83.33	75.29
fa	64.41	50.15	78.37	67.35	99.09	80.43

Table 3: Contribution of layers 1, 1+2, 1+3 for POS tagging task. Higher values in the hierarchy are highlighted.

	Layer 1		*Layer 1+2*		*Layer 1+2+3*	
	P	GP	P	GP	P	GP
en	36.47	36.86	53.68	52.79	53.68	56.81
fa	19.32	29.25	47.48	51.07	61.79	64.50

Table 4: Contribution of layers 1, 1+2, 1+3 for Syntactic Ancestors’ prediction. Higher values in the hierarchy are highlighted.

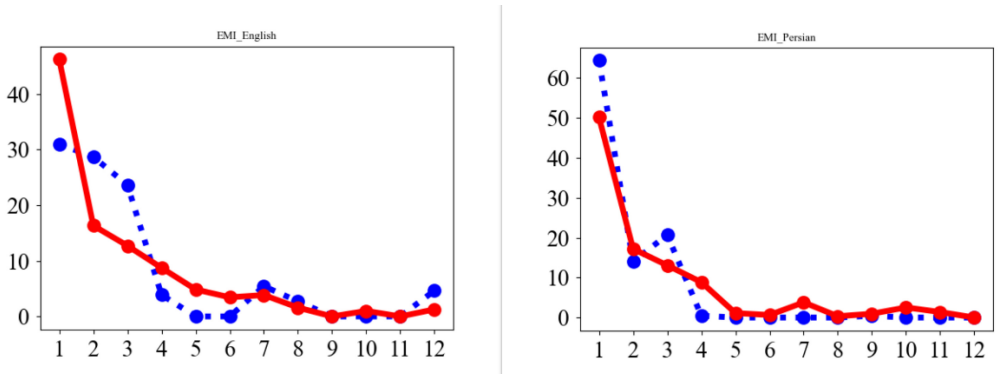

Figure 5: EMI scores for English vs Persian – blue for “MFT” ; red for “nonMFT”

6. Conclusion

After employing the corresponding metrics and implementing probing tasks for English and Persian BERT models, we observed that on some levels there are similarities between the pattern of information emergence and global scores. However, there are substantial differences between the resulting distributions in the two languages that cannot be ignored. These differences highlight the unique characteristics of each language and the way they are captured by the BERT models. It also underscores the importance of considering language-specific factors when interpreting probing results and designing downstream applications.

7. Future Work

The findings of this study open up several avenues for future research. One potential direction is to extend the probing tasks to increase the generality and robustness of our findings. It would also be interesting to explore the impact of different training data or pre-training objectives on the information distribution in the models.

Another promising direction is to refine the probing metrics to capture more nuanced aspects of information emergence, such as the interaction between layers or the temporal dynamics of information processing. This could provide deeper insights into the inner workings of both neural language models and their ability to represent linguistic knowledge.

Finally, the substantial differences observed between English and Persian call for a more thorough analysis of the factors that contribute to these differences. Understanding these factors could help improve the interpretability of neural language models across different languages.

References

1. Guo, Zishan, et al. "Evaluating large language models: A comprehensive survey." *arXiv preprint arXiv:2310.19736* (2023).
2. Naveed, Humza, et al. "A comprehensive overview of large language models." *arXiv preprint arXiv:2307.06435* (2023).
3. Chen, Honghua, and Nai Ding. "Probing the Creativity of Large Language Models: Can models produce divergent semantic association?." *arXiv preprint arXiv:2310.11158* (2023).
4. Blevins, Terra, Omer Levy, and Luke Zettlemoyer. "Deep RNNs encode soft hierarchical syntax." *arXiv preprint arXiv:1805.04218* (2018).
5. Peters, Matthew, et al. “Deep contextualized word representations.” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, https://doi.org/10.18653/v1/n18-1202.
6. Tenney, Ian, Dipanjan Das, and Ellie Pavlick. "BERT rediscovers the classical NLP pipeline." *arXiv preprint arXiv:1905.05950* (2019).
7. Kunz, Jenny, and Marco Kuhlmann. "Where Does Linguistic Information Emerge in Neural Language Models? Measuring Gains and Contributions across Layers." *Proceedings of the 29th International Conference on Computational Linguistics*. 2022.
8. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
9. Farahani, Mehrdad, et al. "Parsbert: Transformer-based model for persian language understanding." *Neural Processing Letters* 53 (2021): 3831-3847.
10. Wolf, Thomas, et al. "Huggingface's transformers: State-of-the-art natural language processing." *arXiv preprint arXiv:1910.03771* (2019).
11. Abadi, Martín, et al. "{TensorFlow}: a system for {Large-Scale} machine learning." *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 2016.
12. Turc, Iulia, et al. "Well-read students learn better: On the importance of pre-training compact models." *arXiv preprint arXiv:1908.08962* (2019).