Probing Persian Language Models

Mahrokh Mirani
Tehran Institute for Advanced Studies
m.mirani@khatam.ac.ir

Faezeh Labbaf
Tehran Institute for Advanced Studies
f.labbaf@khatam.ac.ir

Abstract

Natural language processing has taken big steps Since the advent of Transformers. Stronger language models have been developed that have a great performance in language processing tasks. However, most of these models are trained on English texts and resources and other languages are mostly left to multilingual models. Narrowing our attention to Farsi (Persian) language processing tasks, fortunately, there have been some recent efforts to train famous models on Farsi datasets. In this project, we study some of these models alongside multilingual models to evaluate their understanding of Farsi words and tasks.

1 Introduction

The power of attention mechanism and transformers has completely changed the route of language processing. The increasing performance of language models during recent years has raised many questions about the reasons behind their success. Many researchers are trying to probe different aspects of famous models to find out their performance.

(Rogers et al., 2019) has shown that the first layers of transformer models mostly capture lexical features and more deep layers are responsible for semantic tasks. For more complicated tasks, like reasoning tasks, it remains unknown how and to what extent the models handle implicit information. The Olmpics paper (Talmor et al., 2020) has investigated some state-of-the-art models to evaluate their abilities in difficult symbolic reasoning tasks.

Most of the works in NLP concentrate on the English language, leaving other languages behind. Especially for Farsi, since there are very few works and most of them did not get very good results, it is important to know how these models are working. We examined these models with some reasoning tasks like number or size comprehension, to see how they do. It is important to know if it is even necessary to train strong Farsi models or using multilingual models that have a shared knowledge across different languages is a better approach. Thus we compare existing Farsi models with multilingual BERT as well.

We also perform control experiments to ensure the models' performances depend on their inner knowledge and explain what sense of language do models have.

2 Language Models

Before ParsBERT, all Farsi language models were restricted to RNNs and CNNs. ParsBERT is the first model that trains transformers for Farsi tasks and achieves results that are superior to all of the previous methods (Farahani et al., 2009). It has the same architecture as the bert-base model (Devlin et al., 2019) but is trained on Farsi texts extracted from Wikipedia and other Farsi resources of different categories. We run our experiments and control tasks on ParsBERT to get a better view of its knowledge.

The Hooshvare Lab which released ParsBERT also released other language models for the Farsi language. However, there are no publications or extra explanations on these models and we assume they are trained exactly with the same method as their English counterparts. Among these models, we decided to take RoBERTa-fa and ALBERT-fa for our experiments.

We also conducted our experiments on multilingual BERT which was released along with English BERT and is trained on Wikipedia pages of 104 different languages. One of our objectives is to compare models that are trained on Farsi texts with mBERT which has a comprehensive knowl-

edge obtained from many languages.

3 Probing Tasks

3.1 Comparison

In this section, we examine the ability of language models to compare two numerical values. We examine models on four different tasks, objects size comparison, number comparison, and age comparison numerical and alphabetical. In the objects size comparison, we do not explicitly mention the numerical value, so the model should extract it from its knowledge, despite the other three tasks where the numerical value is given explicitly. We will explain details about each task in the following.

Age And Number Comparison Tasks

To test if our language models can compare two numbers, we conduct three tasks. These tasks are shown in figure 1.

The first task is comparing numerical values of two person's ages; the second task compares two numerical values, and the third task compares alphabetical values of two person's ages. The model needs to encode the numerical or alphabetical values and compare them to handle these tasks. For all three tasks, we used numbers in the range 15 to 38 for the test set and 43 to 105 for the training set, as Talmor et al. (2020) do for their numerical age comparison task to guarantee that model generalizes to values unseen at training time.

Objects' Size Comparison

In this probing task, we examine if models can compare objects' sizes. The format of sentences is shown in figure 1. We replace obj1 and obj2 with objects in the dataset. We translate the dataset from work by Talmor et al. for this task. The dataset contains 35 general objects (like the sun) for the test set and 127 animal objects for the train set.

3.2 Opposition

The goal of this task is to examine how the model recognizes antonyms and synonyms. For this task, we have created a dataset with sentences containing a word and its synonym or antonym. An example of these sentences is shown in figure 1. The mask should be filled with one of two verbs meaning either "was" or "was not". This dataset is gathered using FarsNet (2009) which is a word network for Farsi and split into 2048 train instances and 240 test instances.

3.3 Reasoning

Most of the previous tasks needed some extent of reasoning ability to accomplish. For example, in the object size comparison task, the model needed to understand objects and then compare their sizes. However, to measure the ability of the model in combining its knowledge and inferring new ideas, we have provided a harder reasoning task. The "Breakable" task has two similar formats. In each of them, four objects are thrown from a window and the model has to fill the mask with the one that is more/less probable to break. The dataset is similar to the (Aroca-Ouellette et al., 2021) dataset with some word replacements and translation. The train and test datasets contain 5050 and 710 sentences respectively.

4 Control experiments

In probing the models, the important question is how much knowledge does a model keep in its parameters. Thus we employ two control tasks to evaluate actual knowledge and eliminate effects of training and minor mismatches that happen for each task and model.

4.1 learning curve

In order to measure the effect of fine-tuning on our tasks, we run the tests with different dataset sizes and plotted the accuracies. It is possible that the model can understand all the words in a task and keeps the knowledge necessary for doing it, but its performance is low due to problems in understanding the task itself. In such cases, a small push can help the model increase the accuracy by a meaningful amount. Giving a lot more data can also lead the model to higher performance by teaching it the knowledge it lacks. So we train the model with different datasets of size $N \in \{64, 128, 256, ..., 4096\}$ sliced from train datasets and keep track of accuracy in each case. For each dataset size, a new model is trained with 4 epochs and then tested for accuracy. Since we want to know about the model's knowledge and teaching the model is undesired, we freeze all of the layers of the model except the MLM-head which is optimized to decide the right answer from parameters during training.

Plotting the learning curve for different data

Task	Example	Translation
Age comparison – numerical	«یک فرد num-1 ساله [MASK] از یک فرد num-2 ساله است» الف) بزرگتر ب) کوچکتر	A num-1 years old person is [MASK] than a num-2 years old person. A) older B) younger
Age comparison – alphabetical	«یک فرد num-1 ساله [MASK] از یک فرد num-2 ساله است» الف) بزرگتر ب) کوچکتر	A num-1 years old person is [MASK] than a num-2 years old person. A) older B) younger
Number comparison	«عدد num-1 [MASK] از عدد num-2 است» الف) بزرگتر ب) کوچکتر	The number num-1 is [MASK] than the number num-2 A) bigger B) smaller
Object comparison	«ابعاد یک obj-1 معمولا از ابعاد یک obj-2 بسیار [MASK] است» الف) بزرگتر ب) کوچکتر	The size of a obj-1 is usually [MASK] than the size of a obj-2 A) bigger B) smaller
Antonym-Synonym	«او بسیار adj-1 بود، او adj-2 [MASK].» الف) بود ب) نبود	She/He was very adj-1, she/he [mask] adj-2 A) was B) wasn't
Breakability	«فائزه یک obj-1، یک obj-2، یک obj-3 و یک obj-4 را از پنجره پایین می‌اندازد. احتمال اینکه [MASK] بشکند از بقیه کمتر است". الف)obj-1 ب)obj-2 ج)obj-3 د)obj-4	Faezeh throws a obj-1, a obj-2, a obj-3 and a obj-4 out of the window. The probability that [MASK] breaks is lower than the others A) obj-1 B) obj-2 C) obj-3 D) obj-4

Figure 1: sentence format for different tasks

sizes gives a sense of what the model learns during training. If the learning curve is a straight line with a mild slope, it means that the model is just learning to fill the mask as a new task ignoring its previous knowledge. On the other hand, if there is a jump in the curve it shows that after some training the model has learned how to use its knowledge for the task. We consider two metrics about training: the maximum accuracy that was reached with different data sizes and the weighted sum of all of the data sizes. As we explained, it is preferred if the model has a jump in accuracy with a low amount of data therefore we assign bigger weight to small datasets and decrease the weights as the size grows.

4.2 Language sensitivity

To get a better view of the model's understanding of language, we perform a No-Lang experiment in addition to the original one for each task. In the No-Lang experiment, we want to know how sensitive is the model to language structure and words, so we change the dataset, removing a lot of words in a sentence and keep only the least necessary information for filling the mask. For example, in the size comparison task, the complete sentence "The size of a gorilla is much [MASK] than the size of a salmon" is summarized into "gorilla [MASK] salmon". We also replace answer choices with some random tokens without any meaning in this context in order to prevent the model from getting any sense about the task. With this variation, we try to find out how much the performance depends on the whole sentence. We then plot the learning curve for this task and the original task and compare them. If the two curves are close to each other, that means the model is not using language components efficiently for filling the mask.

5 Implementation Details

All of our models are masked language models that take as input a sequence of tokens and produce a probability distribution over their vocabularies as output. The input sequences are limited to a length of 64 and zero-padded and fed into the model. Then we take the output logits corresponding to mask choices and run a softmax function on them to get their probabilities as final output. For training, all parts of the model are frozen except the mlm-head layer and training lasts for 4 epochs with a learning rate of 2e-5. We used Huggingface libraries (Wolf et al., 2020) for loading and working with datasets and pre-trained models.

Table 1: numerical age comparison results. Accuracy over two answer candidates (random is 50%)

model	Zero Shot	Fine-tune		Lang-Sense
		WS	MAX	
ParsBERT	49.49	45.61	49.72	2.49
ALBERT-fa	49.41	48.03	51.52	0
Multi-lingual BERT	48.9	50.3	57.4	4.5

Table 2: numbers comparison results. Accuracy over two answer candidates (random is 50%)

model	Zero Shot	Fine-tune		Lang-Sense
		WS	MAX	
ParsBERT	49.88	48.44	50.35	0
ALBERT-fa	54.2	57.48	64.33	0.32
Multi-lingual BERT	49.41	50.04	57.57	0

6 Experiments and Analysis

6.1 Results for Number and Age Comparison

Results for numerical age comparison are shown in table 1. As shown, all the models we examine start around random performance in zero-shot and do not obtain any significant progress during the fine-tunning. Also, all the models do not show any language sensitivity. So, we can conclude that our examined models do not encode numbers well in the context of age.

To be sure about the lack of ability to compare

Table 3: Alphabetical age comparison results. Accuracy over two answer candidates (random is 50%)

model	Zero Shot	Fine-tune		Lang-Sense
		WS	MAX	
ParsBERT	49.41	42.49	48.75	1.48
ALBERT-fa	58.1	59.11	61.05	16.13
Multi-lingual BERT	48.9	55.15	60.27	11.2

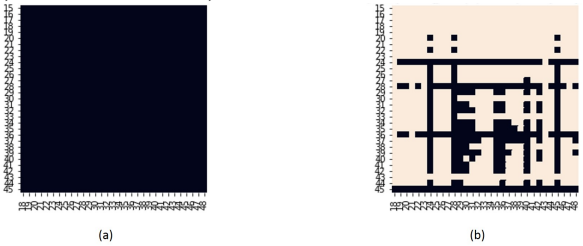
numbers in our models, we conduct the second experiment, which results are shown in the table 2. We conclude that our models do not encode either in age context or in pure number comparison context. Then we examine the alphabetic value of the numbers in the age comparison task to check if models can make the comparison. The results in table 3 show that RoBERTa-fa reaches 70.5 percent accuracy after fine-tunning. Also, the language sensitivity is 16.13 percent, which is greater than other examinations, suggesting that RoBERTa-fa encodes alphabetical values of numbers quite well.

In the following, we will visualize the output of models on the age and number comparison to compare models' outputs with each other. In all the visualizations, we replace "age1" (or "num1") with a number in the range 15 to 35 and "age2" (or "num2") with a number in the range 18 to 38.

Figure 2 shows the output visualization of ParsBERT and ALBERT-fa in the alphabetical age comparison. As shown, ParsBERT always predicts "bigger" for comparison tasks. ALBERT-fa has different output with the majority of predicting "smaller."

Figure 3 Compares output of number comparison for mBERT on English and Farsi input. It shows how mBERT works much more accurately on English input rather than Farsi output. We assume that this is because of the richer resources of the English Wikipedia rather than Farsi Wikipedia.

Figure 2: alphabetical age comparison for a)ParsBERT and b)ALBERT-fa



In conclusion, we saw that Farsi models do not encode numerical values very well, but some of them, like ALBERT-fa, encode alphabetical values of the numbers. Also, it seems that mBERT is encoding the alphabetical values of the numbers as well as ALBERT-fa.

Figure 3: mBERT age comparison for a) English b) Farsi numbers

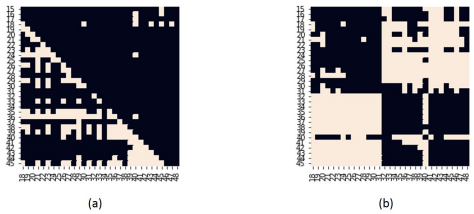


Table 4: Object size comparison results. Accuracy over two answer candidates (random is 50%)

model	Zero-Shot	Fine-tune		Lang-Sense
		WS	MAX	
ParsBERT	37.52	41	54.5	0
ALBERT-fa	47.21	50.8	58	0.33
Multi-lingual BERT	49.41	49.94	54.29	0.68

6.2 Results for Size Comparison

We conduct the experiments on three models, ParsBERT, ALBERT-fa, and Multi-lingual BERT. The results of these experiments are shown in 4.

All the models we examine, start with less than random performance and do not obtain any significant progress during the fine-tunning. Besides, none of the models show language sensitivity, suggesting that the ability to compare object sizes is not encoded in them.

According to the results on English language models shown in (Talmor et al., 2020), BERT-B -which has the same structure as ParsBERT- also does not do well in this task. However, bigger models like RoBERTa-L do a better job, but RoBERTa-L is not trained for Farsi. In fact, the object comparison task needs both knowledge of the numeric value of the object size and the ability to compare. We assume that ParsBERT and BERT-B lack this knowledge due to their smaller structure.

6.3 Results for Opposition

This task is conducted on four Farsi language models Multi-lingual BERT, RoBERTa-fa, ALBERT-fa, and ParsBERT. According to the results shown in table 5. All the models start from random performance, but they progress during

Table 5: Opposition results. Accuracy over two answer candidates (random is 50%)

model	Zero Shot	Fine-tune		Lang-Sense
		WS	MAX	
ParsBERT	45.32	52.98	69.45	0
RoBERTa-fa	44.92	50.21	64.23	0.6
ALBERT-fa	47.65	53.75	66.44	0.89
Multi-lingual BERT	51.91	57.75	76.17	0

fine-tuning and reach a max accuracy of 76 percent. This may lead to the conclusion that our examined language models encode the antonym and synonym adjectives. Looking closely at the learning curves in figure 4, we see that nolang learning curves are mostly above the original learning curve; this means that the model is not sensitive to the input language. According to the learning curves in figure 4, we assume that the models can not distinguish between two target words and confuses in selecting one of them. When we replace target words with non-sense words in nolang, the models perform better.

Talmor et al.(2020) also have examined a similar probing task. Most of the models they examined start from random performance and reach around 80 percent during fine-tunning. We believe that it is because the negation word (not) is a whole separate token, and it is easier for language models to detect it.

Figure 4: a)ALBERT-fa, b)mBERT, c)RoBERTa-fa and d)ParsBERT learning curve for Opposition task

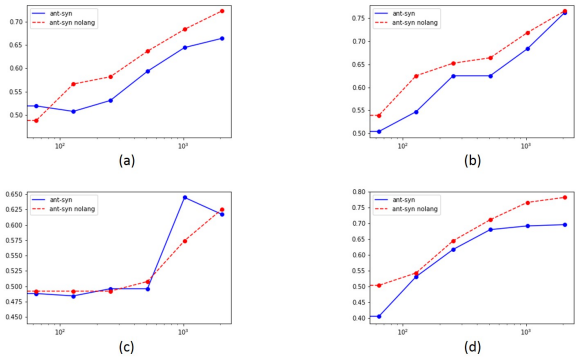


Table 6: Breakability results. Accuracy over four answer candidates (random is 25%)

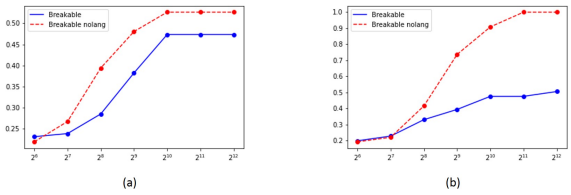
model	Zero Shot	Fine-tune		Lang-Sense
		WS	MAX	
ParsBERT	23.12	32.60	47.35	0.27
RoBERTa-fa	19.83	32.79	50.52	0.29

6.4 Results for breakability

The last task we prob is the breakability task. Models should determine which of the four fallen objects is mostly foolan to break or not break. We prob this task over ParsBERT and RoBERTa-fa to compare the results with the similar work that Aroca-Ouellette et al.(2021) did before. Our results are shown in table 6. Based on the results, both models start from lower performance than random, but very soon, the performance increases in both models. However, this observation is not sufficient since both models' language sensitivity is very low. Looking at models learning curves in 5, we understand that this progress in performance is not obtained by models' knowledge but by cheating from the train set.

It shows that none of these models can determine if an object is breakable; they just learn to classify the objects into two classes and choose the correct answer based on these classes.

Figure 5: a)RoBERTa and b)BERT learning curve for breakable task



7 Conclusion

In this work, we prob Farsi language models on six different tasks and report the results. The overall results were not good and showed that models trained on Farsi texts are still weak and need more effort to reach an acceptable level. One possible reason is that the models are not strong enough yet. For example, ParsBERT is BERT-base, trained on Farsi data. While BERT-base it-

self is not a very strong model and performs worse on similar tasks than BERT-large. Another reason may be the training data. The main data for training is language models is Wikipedia, while Farsi Wikipedia is not that rich to answer the language models.

References

Talmor, Alon, Yanai Elazar, Yoav Goldberg and Jonathan Berant. 2020. *oLMpics-On What Language Model Pre-training Captures.* Transactions of the Association for Computational Linguistics, 8, 743-758.

Stephane Aroca-Ouellette, Cory Paik, Alessandro Roncone and Katharina Kann. 2021. *PROST: Physical Reasoning of Objects through Space and Time* ArXiv, abs/2106.03634.

Rouhizadeh, M. 2009. *Developing the Persian WordNet of Verbs ; Issues of Compound Verbs and Building the Editor.*

Farahani, M., Gharachorloo, M., Farahani, M., Manthouri, M. 2020. *ParsBERT: Transformer-based Model for Persian Language Understanding.* ArXiv, abs/2005.12515.

Devlin, J., Chang, M., Lee, K., Toutanova, K. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* NAACL-HLT.

Rogers, A., Kovaleva, O., Rumshisky, A. 2021. *A Primer in BERTology: What We Know About How BERT Works.* Transactions of the Association for Computational Linguistics, 8, 842-866.

Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and Rémi Louf and Morgan Funtowicz and Joe Davison and Sam Shleifer and Patrick von Platen and Clara Ma and Yacine Jernite and Julien Plu and Canwen Xu and Teven Le Scao and Sylvain Gugger and Mariama Drame and Quentin Lhoest and Alexander M. Rush 2020. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*