Attention Understands Semantic Relations

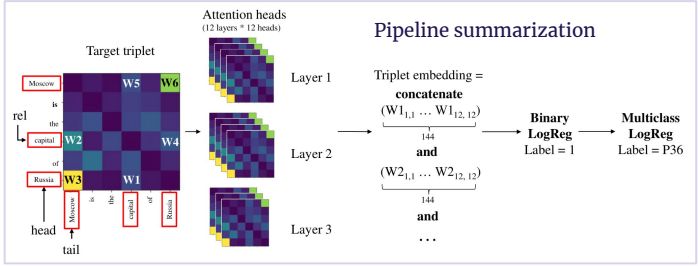
Anastasia Chizhikova, Sanzhar Murzakhmetov, Oleg Serikov, Tatiana Shavrina, Mikhail Burtsev

Introduction

LMs are used in almost every NLP tasks, yet they are often criticized for *memorizing* instead of *generalizing* knowledge. As a result, model interpretation research area has been actively developing. Probing papers mostly focus on investigating grammatical relations, while semantic knowledge remains less studied.

To narrow the gap, we propose a simple Relation Extraction (RE) approach and interpret our outcomes.

We perform on and analyze the behaviour of the **BERT self-attentions mechanisms.**



Data and Methodology

Data: subset of the **TRex** dataset with texts annotated for relation triplets of 95 semantic relation types from WikiData.

RE Pipeline:

1) collect a vector of attention weights for each triplet from every attention map in the model
2) **binary classifier** identifies if a triplet has a semantic relation between its tokes
3) **multiclass classifier** labels meaningful triplets with a relation Id

-> the output is our **RE model prediction**

Interpretation:

1) Performance of the binary classifier -> **Is a simple linear model capable of generalizing over a knowledge encoded in the attention mechanism?**

2) Performance of the multiclass model -> **How informative attention weights are in terms of their capability to differentiate semantically close relation types?**

3) The effect of different weights in the attention matrix and different layers on the result -> **How is the knowledge distributed across the attention mechanism?**

4) Multiclass classifiers' features importances -> **Are similar relations encoded similarly?**

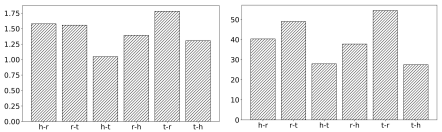
Results

Model	Dataset	Pr	Rec	F1
Our	T-REx	0.220	0.534	**0.312**
REBEL	T-REx	0.223	0.375	0.276
Our	DocRED	0.259	0.175	0.208
REBEL	DocRED	0.594	0.437	**0.503**

Performance of our RE pipeline compared with the REBEL (Cabot and Navigli, 2021) performance on two datasets

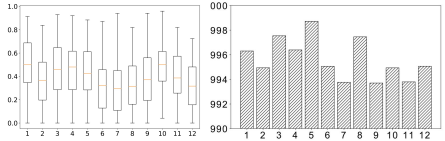
Attention	Classifier	Pr	Rec	F1	Acc
BERT	Binary				0.902
Random BERT	Binary				0.498
BERT	Multi	0.867	0.861	0.863	
Random BERT	Multi	0.026	0.024	0.018	

Classifiers performance on a test set compared with a randomly initialised BERT. Probes are selective.



Feature importance values averaged by the attention type for the binary (left) and multiclass (right) classifier. None of the weights can be neglected in the RE task, as there is no dramatic difference between them.

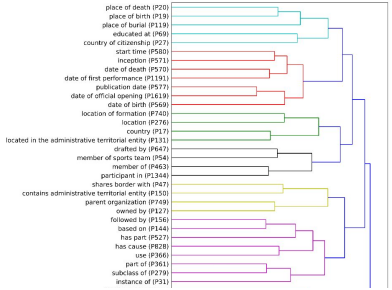
Relation prediction scores on the weights of only one layer of the self-attention mechanism (left) and one layer of the BERT layers embeddings (right). The distribution of knowledge appears to be similar.



Clusterization

The results of agglomerative clustering of the feature importance weights with cosine distance shows that the relation types are logically organised into semantic groups. A complicated way of how knowledge is structured inside language models is not purely stochastic but can be interpreted.

We find no strict mapping between the attention heads and the semantic relation types.



Conclusion

We introduce a novel approach to interpreting Language models and use it to study BERT's awareness of semantic relations

We find that:

- semantic relations of different types are encoded with a combination of attention weights provided by different heads
- attention weights are not as informative as layers' units activations but provide a reliable, straightforward approach to ranking the layers' awareness of relational linguistic features
- none of the layers and attentions must be neglected while developing an unsupervised approach to relation extraction
- there are no individual relation-specific heads, yet one could meaningfully group relations by the heads' relevance for them
- graphs are available through the link via qr code