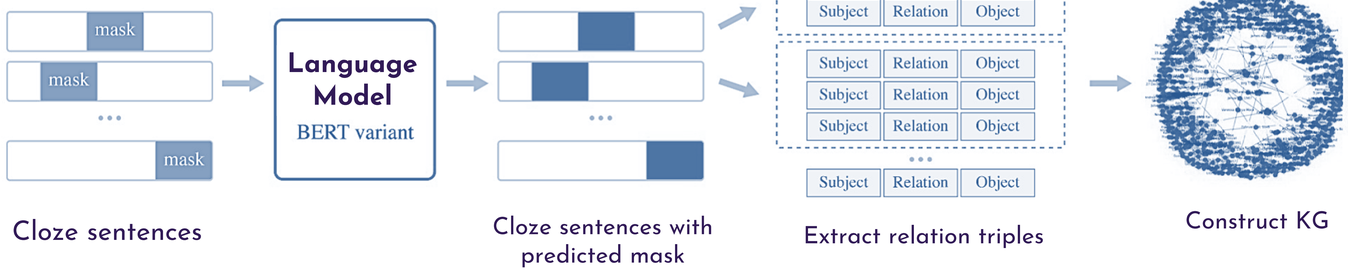
Interpreting Language Models Through Knowledge Graph Extraction

VINITRA SWAMY, ANGELIKA ROMANOU, MARTIN JAGGI

{firstname.lastname@epfl.ch}

Knowledge Graph Extraction Pipeline



Models

- Distill BERT
- BERT
- RoBERTa + trained from scratch

Pretrained models are from the HuggingFace Library

Datasets

- SQuAD
- Google-RE: Place of Birth, Date of Birth, Place of Death

Datasets are inspired from Facebook's LAMA paper

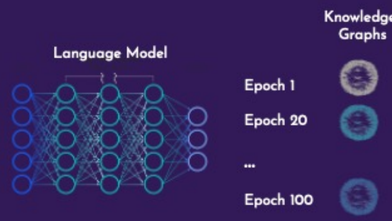
GitHub Repository

github.com/epfml/interpret-lm-knowledge

Motivation

How can we diagnose strengths and weaknesses of transformer-based language models beyond traditional accuracy metrics?

We extract snapshots of acquired knowledge at sequential stages of the training process.



Knowledge acquisition timeline: what does the language model learn at what point in time?

Research Questions

1. *Quantitatively compare knowledge acquisition across language models*
2. *Analyze the same model at different stages over time (early training)*
3. *Compare knowledge graphs linguistically*

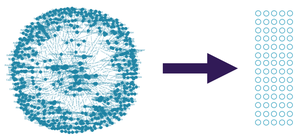
Knowledge Graph Metrics

Graph-Edit-Distance

$$GED(g_1, g_2) = \min_{(e_1,\ldots,e_k)\in\mathcal{P}(g_1,g_2)} \sum_{i=1}^{k} c(e_i)$$

Graph2Vec

+ Euclidean Distance



These metrics are inspired from graph literature to quantitatively compare KG extracts.

Linguistic Metrics

Part-of-Speech Overprediction Rate

$$POSOR(pos) = \frac{(LM_{pos} - GLM_{pos}) \cdot 100}{GLM_{pos}}$$

POSOR allows us to identify part-of-speech deficiencies of our language models. This metric is framed within the context of probing task literature.

Probing Tasks

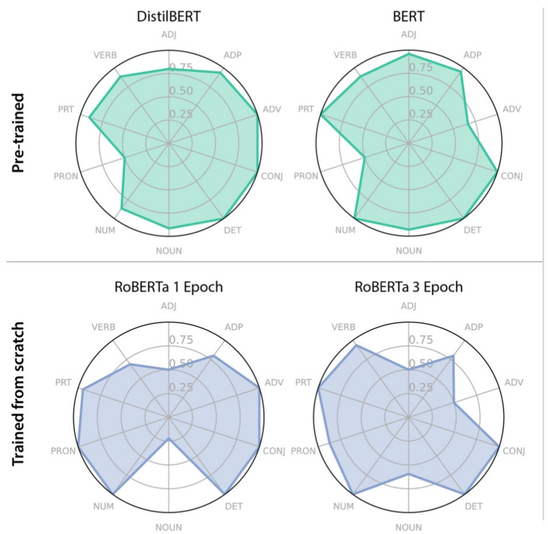
- wh- words
- prepositions
- coordinate conjunctions
- negation
- coordination
- EOS
- spatial

Knowledge Graph Results

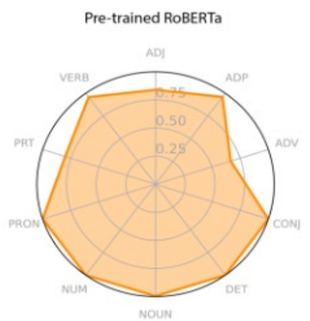
Target Model (distance from RoBERTa)	Graph-Edit-Distance on the extracted knowledge graph	Euclidean distance on the graph2vec embeddings
RoBERTa 1e	141.25	0.2260
RoBERTa 3e	135.00	0.1733
RoBERTa 5e	130.50	0.1607
RoBERTa 7e	121.50	0.1605
DistilBERT	28.50	0.0284
BERT	16.50	0.0202

Trained-from-scratch ↓ / Pretrained ↓

Across both quantitative graph metrics, we see the distance from each model to pretrained RoBERTa reduce as the number of epochs and the amount of training data increase.



Linguistic Results



These radio graphs visualize the POSOR metric across different models and epochs. By learning model part-of-speech limitations, we can answer the question: what kind of data should I train my model on to improve it further?

NEURAL INFORMATION PROCESSING SYSTEMS

EPFL

December 2021