# BERT Rediscovers the Classical NLP Pipeline

Flavjo Xhelollari

# What’s going on?

The authors’ claims and aims:

1. Within a neural network, where is the linguistic information captured?
2. The model has (localizable) steps of a traditional NLP pipeline.
3. Order [of pipelines] is: POS tagging, parsing, NER, Semantic Roles (srl) and coreference.
4. *The model adjusts the pipeline dynamically: revise low-level decision on basis of “making clear” that information is represented from higher levels.

# Procedure

- * Do these models really learn the “abstractions” we think make a language up? Is it just a complex statistical procedure that deals with co-occurrences?
- Previous research [via reverse-engineering] shows that deep LM encode a variety of worthy info: the more complex structures are represented in higher hierarchies of the model.
- Authors try to find where this happens. They introduce two new findings:

1. Order of encoding of abstractions resembles traditional NLP pipeline.
2. Processing individual sentences (layer-by-layer) by the BERT network.

# Approach

- Use Edge Probing :  how well information about linguistic structure can be extracted from a pre-trained encoder. Have access only to per-token context vector, i.e. in a token basis.
- BERT - replace ELMo’s bidirectional LSTM encoding with transformers. Focus on next-sentence predictions. Use ‘masked SA’ (random words void) and word piece tokenization (e.g. walking = walk + ing - 2 tokens).
- Two metrics used: 1.Scalar Mixing Weights used to pool across layers, and shows the “relevance” of info related to task. 2. Center of gravity as a summary of statistics.
- The authors aim to estimate at which layer in the encoder a target can be accurately predicted. They train a series of cumulative classifiers using scalar mixing.

# Results

- Observed a trend: POS tags processed earliest, followed by constituents, dependencies, semantic roles, and coreference, i.e. basic info appears early.
- Syntax is more localized in a few layers, while semantics are spread throughout the network.
- Most examples are classified early due to heuristic shortcuts.
- The same general ordering on the 12-layer BERT-base model. The representations for a given task tend to concentrate at the same layers relative to the top of the model.
- Conclusion: The network can sometimes resolve structure out-of-order by using high-level information. This supports the idea that deep language models can handle the necessary abstractions for language processing and complex interactions between levels of information.