Exploring numerical representations of language units

Tomasz Limisiewicz

24 September 2020
Introduction
In a Nutshell

• What do I examine: Internal representation of Neural Networks Contextual word representation, mainly language models: BERT, GPT, XLM

• Aims of exploration: Examining learned linguistic features, improving transfer between tasks and languages.

• Examined Neural Networks: Mainly contextual representation of language models (BERT, GPT, XLM).
Methods of exploration

• Structural analysis: Answers the question: "Does particular components of the Neural Network know something about specific linguistic properties"?

• Behavioral analysis: Make inference of the models internal representation based on its behaviour in particular cases.
• A survey of structural analysis: Belinkov and Glass 2019

• A survey of syntax representation in Neural Networks and Word Embeddings: Limisiewicz and Mareček 2020
My Points of Focus

- Explanation of Neural Network representation
- Multilingual approaches, going beyond English
- Separation of task-specific information
Probing
• Contextual neural network models is trained, e.g. for Language Modeling, Translation
• The parameters of the network are fixed (frozen). A new simple network takes is trained on top for auxiliary linguistic task, e.g. POS tags prediction.
• We assume that when probing classifier accuracy is high the networks encodes linguistic abstraction well.

Liu et al. (2019): “Linguistic Knowledge and Transferability of Contextual Representations”
Figure 1: Accuracy of POS tag probing from RNN latent vectors compared with static word embeddings
Figure 2: Accuracy of POS tag probing from RNN representation by the pre-training objective
Figure 3: Comparison of two widely used syntactic structure types: dependency and constituency trees, from Jurafsky and Martin 2009
Hewitt and Manning 2019 multiply contextual vectors by trainable matrix to approximate syntactic tree distance between tokens by the L2 norm of the difference of the transformed vectors.

\[
\min_B \left| (B(h_i - h_j))^T (B(h_i - h_j)) - d_T(w_i, w_j) \right|
\] (1)

This approach produces the approximate syntactic pairwise distances for each pair of tokens. The minimum spanning tree is used to create a dependency tree with high accuracy (82.5% UAS on Penn Treebank).
Orthogonal Probing
Orthogonal Probing
Orthogonal Probing

Introduction

Probing

Attention matrices

Challenges
Orthogonal Probing

Introduction

Probing

Attention matrices

Challenges
Orthogonal Probing

Introduction

Probing

Attention matrices

Challenges
Figure 4: Lexical and syntactic information is separated by a syntactic probe
Advantages and Issues

+ Good results in induction of the Syntactic Trees
+ Clearly visible distinction between layer
+ Possibility to reduce number of dimension of the vector
  − Strong supervision of the probe
  − The structure can be memorized in the additional layer instead of being encoded in the representation
Attention matrices
there is considerable energy saving potential in public buildings, for example, which would facilitate the transition towards a stable, green economy.

**Figure 5**: Self-attention in a particular heads of a language model aligns with dependency relations: adjective modifier, objectives.
Advantages and Issues

+ Less supervision needed, observations based on qualitative analysis
+ Is not restricted by annotation guidelines
  – Some annotation is needed to automatically identify syntactic heads
  – Generally gives worse results than structural probing
Challenges
• Chi, Hewitt, and Manning 2020 probed multilingual model for syntactic structure. They evaluated transfer between languages.
• Kulmizev et al. 2020 probed and compared two annotation styles (UD and SUD).
• Limisiewicz, Rosa, and Mareček 2020 we extracted syntactic trees from multilingual model.

All the approaches used multilingual BERT.
Multi-tasking

- Syntax
- Lexicology
- Discourse: coreferences
- Semmantics
- ...

Introduction  Probing  Attention matrices  Challenges