

LSD: Linguistic Structure Representation in Neural Networks

David Mareček, Jindřich Libovický, Rudolf Rosa, Tomáš Musil

🖬 September 17, 2019



Charles University Faculty of Mathematics and Physics Institute of Formal and Applied Linguistics



Outline

Intro and Motivation

Word Embeddings and Morphology

Multi-head Self-attentions and Syntax

References

Intro and Motivation

- Word embeddings and Deep neural networks perform great
- They do not have any explicit knowledge of linguistic abstractions
- How do they work? What emergent abstractions can we observe in them? How can we interpret them?
- Are the emergent structures and abstractions similar to classical linguistic structures and abstractions?

LSD project

 $\label{eq:Linguistic Structure representation in Deep networks$

- National Science Foundation of Czech Republic
- 2018 2020



Goals:

- Word embeddings and DNNs perform great.
- They do not have any explicit knowledge linguistic abstractions.
- How do they work? What abstractions can we observe in them? How do we interpret them?
- Are the emergent structures similar to classical linguistic structures?

LSD team



Selected LSD results

Inspecting Word Embeddings using Principal Component Analysis (Musil, 2019)

• What features are important for word embeddings of various NLP tasks?

Derivational Morphological Relations in Word Embeddings (Musil et al., 2019)

• Unsupervised clustering of word-embedding differences captures derivational relations.

Neural Networks as Explicit Word-Based Rules (Libovický, 2019)

• We interpret a convolutional network for sentiment classification as word-based rules.

Looking for Syntax in Transformer Self-Attentions (Mareček and Rosa, 2019, 2018)

Building constituency trees from multi-head self-attentions.

Word Embeddings and Morphology

- A vector for each word (e.g. 100 dimensional, i.e. each word associated with a list of 100 real numbers)
- · Learned in an unsupervised way from large plaintext corpora
- Observes the distributional hypothesis: words that appear in similar context have similar embeddings

Principal Component Analysis (PCA)

- Transformation to another orthogonal basis set
- 1st principal component has the largest possible variance across the data
- Each other principal component is orthogonal to all preceding components and has the largest possible variance.
- If something correlates with the highest principal components its possibly very important for the NLP task.



Independent Variable x

Word-embeddings learned by NMT, correlation with POS tags













What is the separated island of Nouns visible in PCA2?

When we take a sample of words from this cluster, it contains almost exclusively named entities:

Fang, Eliáš, Još, Aenea, Bush, Eddie, Zlatoluna, Gordon, Bellondová, Hermiona

Word-embedding space learnt by Sentiment Analysis

- Task: deciding whether a given text is emotionally positive, negative, or neutral.
- Trained on Czech ČSFD database (https://www.csfd.cz/), data were obtained from user comments and rankings of movies.
- Architecture: Convolutional neural network based on Kim (2014).

Neg: "Very boring. I felt asleep." **Pos:** "Great movie with super effects!!!"



Word-embedding space learnt by Sentiment Analysis



We sampled some words from the vector space...

Word-embedding space learnt by Sentiment Analysis



Looking for derivational relations

e.g. kompenzovat – kompenzace (compensate – compensation)



Looking for ???

• Future work: meaning?

Multi-head Self-attentions and Syntax

Transformer NMT



Multi-headed Self-attention Mechanism

- Encoder: 16 attention heads \times 6 layers
- Attention = weighted sum of "all" "words"
 - "word": contextual representation of word position from previous layer
 - "all": usually focused on just one word



Source: Attention is all you need (Vaswani et al., 2017)

Observation

- Visualisation: matrix of attention weights
- Common pattern: balustrade
 - Baluster: continuous sequence of words attending to the same position
 - Looks like a syntactic phrase
 - Usually attends to phrase boundary
- Research questions
 - Is that syntactic?
 - To what extent?
 - What kind of syntax?



Balustrades (70% of the attention heads)



Diagonals (especially 1st layer)



Attend to end, mixed, scattered...



Approach

- 1. Transformer NMT: French \leftrightarrow English, German \leftrightarrow English, French \leftrightarrow German
- 2. Balusters \rightarrow phrase candidates
 - Phrase score: average attention weight (summed and equalized)
- 3. Phrase candidates \rightarrow binary constituency tree
 - Linguistically uninformed algorithm
 - Tree score = sum of phrase scores
 - CKY: find tree with maximal score
- 4. Compare to standard constituency syntactic trees
 - Penn Treebank, French Treebank, Negra Corpus (via Stanford parser)
 - we observe a 40% match
 - baseline has a 30% match (right-aligned balanced binary tree)

Results



- The emergent structures can be seen as syntactic to some extent
- Shorter phrases are often captured
- Sentence clauses are often captured

LSD: Linguistic Structure Representation in Neural Networks

Summary

- 1. Word embeddings and neural networks are trained without linguistic information.
- 2. Examine emergent abstractions and structures!
- 3. Some morphology is captured.
- 4. Some syntax is captured.

https://ufal.cz/grants/lsd



- Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the* 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, 2014. doi: 10.3115/v1/d14-1181. URL
 - http://dx.doi.org/10.3115/v1/d14-1181.
- Jindrich Libovický. Neural networks as explicit word-based rules. *CoRR*, abs/1907.04613, 2019. URL http://arxiv.org/abs/1907.04613.
- David Mareček and Rudolf Rosa. Extracting syntactic trees from transformer encoder self-attentions. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 347–349, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5444. URL https://www.aclweb.org/anthology/W18-5444.
- David Mareček and Rudolf Rosa. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4827.

Tomáš Musil. Examining structure of word embeddings with pca. In Kamil Ekštein, editor,

Text, Speech, and Dialogue, pages 211–223, Cham, 2019. Springer International Publishing. ISBN 978-3-030-27947-9.

Tomáš Musil, Jonáš Vidra, and David Mareček. Derivational morphological relations in word embeddings. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 173–180, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-4818.