A brief survey on the Structural Probe paper by John Hewitt and Christopher Manning

Paul-André Melliès

Institut de Recherche en Informatique Fondamentale (IRIF)

CNRS – Université Paris Cité – INRIA

Panel Discussion

AITP — 4 September 2024

The paper

A Structural Probe for Finding Syntax in Word Representations

published in 2019

by Christopher Manning and his PhD student John Hewitt

A Structural Probe for Finding Syntax in Word Representations

John Hewitt
Stanford University
johnhew@stanford.edu

Christopher D. Manning
Stanford University
manning@stanford.edu

Abstract

Recent work has improved our ability to detect linguistic knowledge in word representations. However, current methods for detecting syntactic knowledge do not test whether syntax trees are represented in their entirety. In this work, we propose a *structural probe*, which evaluates whether syntax trees are embedded in a linear transformation of a neural network's word representation space. The probe identifies a linear transformation under which squared L2 distance encodes the distance between words in the parse tree, and one in which squared L2 norm encodes depth in the parse tree. Using our probe, we show that such transformations exist for both ELMo and BERT but not in baselines, providing evidence that entire syntax trees are embedded implicitly in deep models' vector geometry.

1 Introduction

As pretrained deep models that build contextualized representations of language continue to provide gains on NLP benchmarks, understanding what they learn is increasingly important. To this end, probing methods are designed to evaluate the extent to which representations of language encode particular knowledge of interest, like part-of-speech (Belinkov et al., 2017), morphology (Peters et al., 2018a), or sentence length (Adi et al., 2017). Such methods work by specifying a *probe* (Conneau et al., 2018; Hupkes et al., 2018), a supervised model for finding information in a representation.

Of particular interest, both for linguistics and for building better models, is whether deep models' representations encode syntax (Linzen, 2018). Despite recent work (Kuncoro et al., 2018; Peters et al., 2018b; Tenney et al., 2019), open questions remain as to whether deep contextual models encode entire parse trees in their word representations.

In this work, we propose a *structural probe*, a simple model which tests whether syntax trees are consistently embedded in a linear transformation of a neural network's word representation space. Tree structure is embedded if the transformed space has the property that squared L2 distance between two words' vectors corresponds to the number of edges between the words in the parse tree. To reconstruct edge directions, we hypothesize a linear transformation under which the squared L2 norm corresponds to the depth of the word in the parse tree. Our probe uses supervision to find the transformations under which these properties are best approximated for each model. If such transformations exist, they define inner products on the original space under which squared distances and norms encode syntax trees – even though the models being probed were never given trees as input or supervised to reconstruct them. This is a structural property of the word representation space, akin to vector offsets encoding word analogies (Mikolov et al., 2013). Using our probe, we conduct a targeted case study, showing that ELMo (Peters et al., 2018a) and BERT (Devlin et al., 2019) representations embed parse trees with high consistency in contrast to baselines, and in a low-rank space.[1]

In summary, we contribute a simple structural probe for finding syntax in word representations (§2), and experiments providing insights into and examples of how a low-rank transformation recovers parse trees from ELMo and BERT representations (§3,4). Finally, we discuss our probe and limitations in the context of recent work (§5).

2 Methods

Our goal is to design a simple method for testing whether a neural network embeds each sentence's

[1]We release our code at https://github.com/john-hewitt/structural-probes.

Proceedings of NAACL-HLT 2019, pages 4129–4138
Minneapolis, Minnesota, June 2 - June 7, 2019.

General idea of the paper

The idea is to probe large language models for their ability to capture

the Stanford Dependencies formalism

claiming that capturing most aspects of the formalism implies

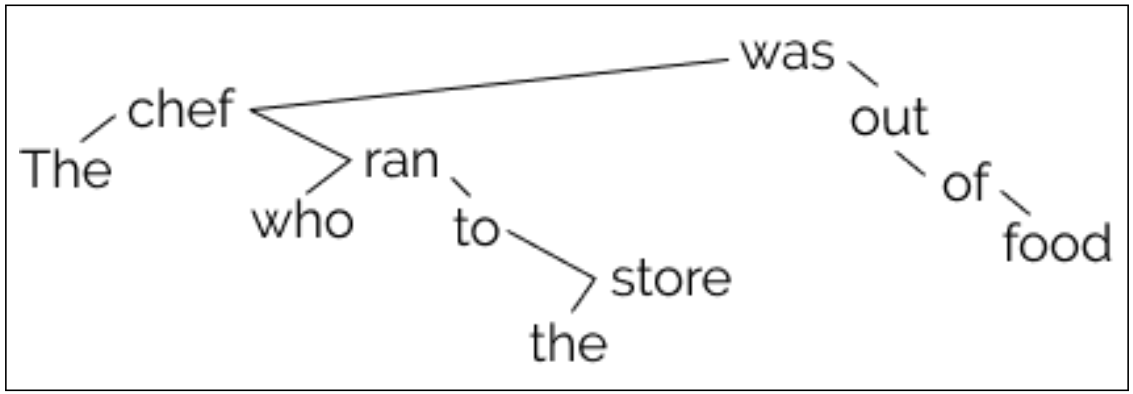
an understanding of English syntactic structure

To this end, the idea is to obtain fixed word representations for sentences of

the parsing train/dev/test splits of the Penn Treebank

with no pre-processing.

Dependency parse tree



An old idea by the linguist Lucien Tesnière (1893-1954) at the heart of

the **Universal Dependencies** project
a framework for consistent annotation
of grammar across different human languages.

The language model

We suppose given a language model $\mathcal{M}$ that takes in

a sequence of n words $w^{\ell}_{1:n}$

and produces

a sequence of vector representations $\mathbf{h}^{\ell}_{1:n}$

where the number ℓ identifies the sentence.

Preliminaries on inner products

Starting with the dot product, we can define a family of inner products,

$$\langle \mathbf{h}, \mathbf{h} \rangle_A \quad = \quad \mathbf{h}^T A \, \mathbf{h}$$

parameterized by any positive semidefinite, symmetric matrix

$$A \in \mathbb{R}_+^{m \times m}$$

The inner product can be equivalently defined using a linear transformation

$$B \in \mathbb{R}^{k \times m}$$

such that $A = B^T B$. One then obtains:

$$\langle \mathbf{h}, \mathbf{h} \rangle_A \quad = \quad B\mathbf{h}^T \cdot B\mathbf{h} \quad = \quad \langle B\mathbf{h}, B\mathbf{h} \rangle$$

Preliminaries on inner products

Every inner product corresponds to a distance metric:

$$\begin{aligned} d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell) &= \langle \mathbf{h}_i^\ell - \mathbf{h}_j^\ell \,,\, \mathbf{h}_i^\ell - \mathbf{h}_j^\ell \rangle_A \\ &= \langle B(\mathbf{h}_i^\ell - \mathbf{h}_j^\ell) \,,\, B(\mathbf{h}_i^\ell - \mathbf{h}_j^\ell) \rangle \end{aligned}$$

where i, j index the word in the sentence.

The structural probe

The parameters of the probe are exactly the matrix B, which is trained

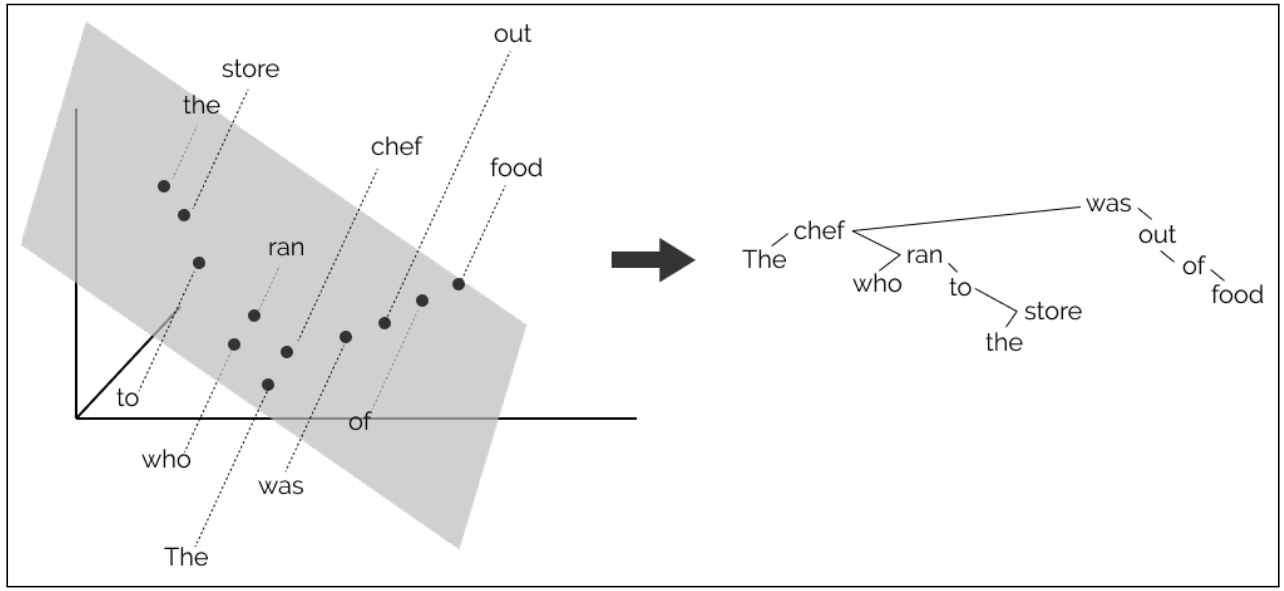
> to **recreate the tree distance** between all pairs of words (w_i^ℓ, w_j^ℓ) in all sentences T^ℓ in the training set of a parsed corpus

Approximated through gradient descent:

$$min_B \sum_\ell \frac{1}{|s^\ell|^2} \left| d_{T^\ell}(w_i^\ell, w_j^\ell) - d_B(\mathbf{h}_i^\ell, \mathbf{h}_j^\ell) \right|$$

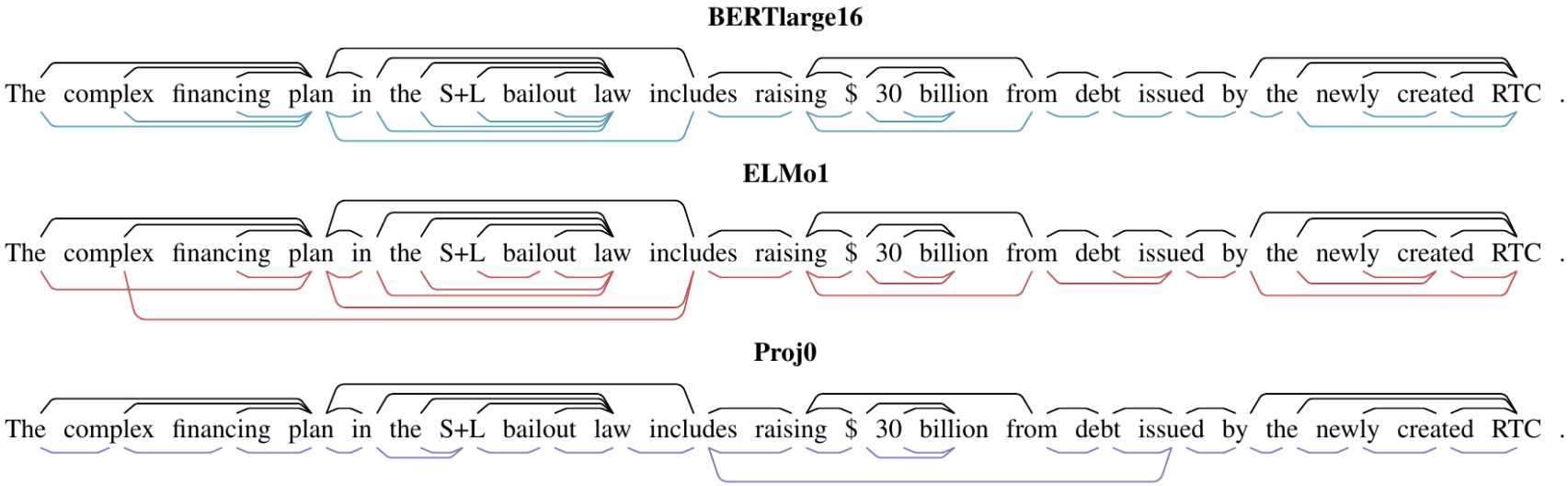
where $|s^\ell|$ is the length of the sentence.

Parse tree and its vectorial representation



Dependency parse trees

Illustration of dependency parse trees obtained by the structural probe:

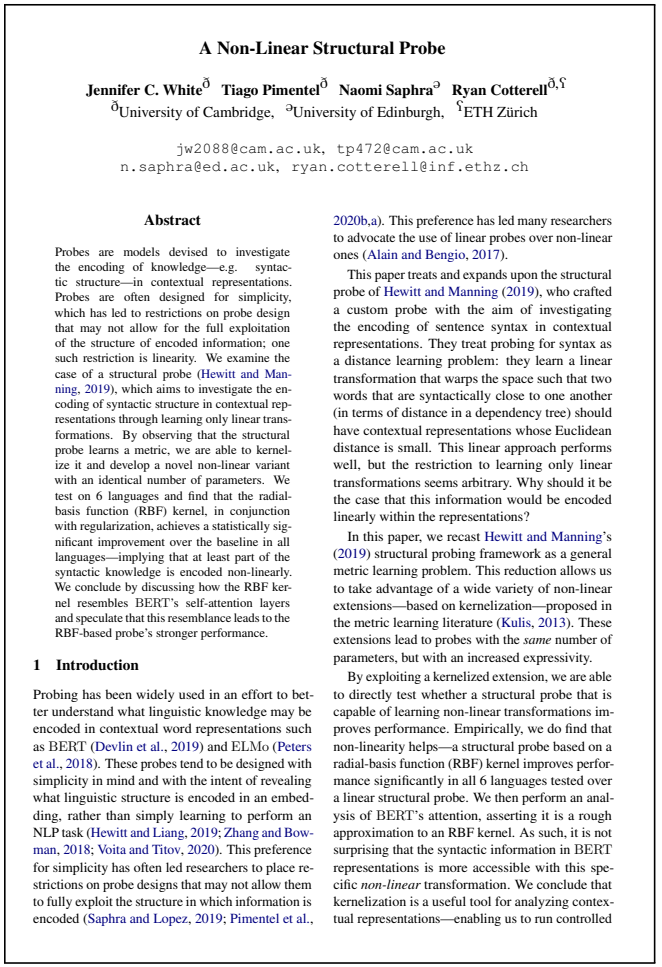


The paper

A Non-Linear Structural Probe

published in 2021

by Jennifer C. White
Tiago Pimentel
Naomi Saphra
Ryan Cotterell



A Non-Linear Structural Probe

Jennifer C. White[ð] **Tiago Pimentel**[ð] **Naomi Saphra**[ə] **Ryan Cotterell**[ð,ʕ]
[ð]University of Cambridge, [ə]University of Edinburgh, [ʕ]ETH Zürich

jw2088@cam.ac.uk, tp472@cam.ac.uk
n.saphra@ed.ac.uk, ryan.cotterell@inf.ethz.ch

Abstract

Probes are models devised to investigate the encoding of knowledge—e.g. syntactic structure—in contextual representations. Probes are often designed for simplicity, which has led to restrictions on probe design that may not allow for the full exploitation of the structure of encoded information; one such restriction is linearity. We examine the case of a structural probe (Hewitt and Manning, 2019), which aims to investigate the encoding of syntactic structure in contextual representations through learning only linear transformations. By observing that the structural probe learns a metric, we are able to kernelize it and develop a novel non-linear variant with an identical number of parameters. We test on 6 languages and find that the radial-basis function (RBF) kernel, in conjunction with regularization, achieves a statistically significant improvement over the baseline in all languages—implying that at least part of the syntactic knowledge is encoded non-linearly. We conclude by discussing how the RBF kernel resembles BERT's self-attention layers and speculate that this resemblance leads to the RBF-based probe's stronger performance.

1 Introduction

Probing has been widely used in an effort to better understand what linguistic knowledge may be encoded in contextual word representations such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018). These probes tend to be designed with simplicity in mind and with the intent of revealing what linguistic structure is encoded in an embedding, rather than simply learning to perform an NLP task (Hewitt and Liang, 2019; Zhang and Bowman, 2018; Voita and Titov, 2020). This preference for simplicity has often led researchers to place restrictions on probe designs that may not allow them to fully exploit the structure in which information is encoded (Saphra and Lopez, 2019; Pimentel et al., 2020b,a). This preference has led many researchers to advocate the use of linear probes over non-linear ones (Alain and Bengio, 2017).

This paper treats and expands upon the structural probe of Hewitt and Manning (2019), who crafted a custom probe with the aim of investigating the encoding of sentence syntax in contextual representations. They treat probing for syntax as a distance learning problem: they learn a linear transformation that warps the space such that two words that are syntactically close to one another (in terms of distance in a dependency tree) should have contextual representations whose Euclidean distance is small. This linear approach performs well, but the restriction to learning only linear transformations seems arbitrary. Why should it be the case that this information would be encoded linearly within the representations?

In this paper, we recast Hewitt and Manning's (2019) structural probing framework as a general metric learning problem. This reduction allows us to take advantage of a wide variety of non-linear extensions—based on kernelization—proposed in the metric learning literature (Kulis, 2013). These extensions lead to probes with the *same* number of parameters, but with an increased expressivity.

By exploiting a kernelized extension, we are able to directly test whether a structural probe that is capable of learning non-linear transformations improves performance. Empirically, we do find that non-linearity helps—a structural probe based on a radial-basis function (RBF) kernel improves performance significantly in all 6 languages tested over a linear structural probe. We then perform an analysis of BERT's attention, asserting it is a rough approximation to an RBF kernel. As such, it is not surprising that the syntactic information in BERT representations is more accessible with this specific *non-linear* transformation. We conclude that kernelization is a useful tool for analyzing contextual representations—enabling us to run controlled

Thank you!