BERT Rediscovers the Classical NLP Pipeline

Ian Tenney[*1], Dipanjan Das[1], and Ellie Pavlick[1,2]*
[1]Google Research, [2]Brown University

Paper

Code

Overview

Question: Does BERT [1] learn linguistic abstractions, or is it just really good at summarizing co-occurrence statistics?

- BERT is a deep model. Do the layers make sequential decisions?
- Is linguistic information localized in different layers of the encoder?

Takeaways:

- Linguistic abstractions appear in a consistent order, with POS tagging in lower layers, followed by parsing, NER, semantic roles, then coreference.
- But, individual decisions don't always follow this: low-level decisions can be revised based on high-level information.

BERT by Layer

For each task τ, train probing classifiers $\{P_\tau^{(\ell)}\}$ for $\ell = 0, 1, \ldots, L$

Scalar Mixing Weights:

ELMo-style: let $s_\tau = \mathrm{softmax}(a_\tau)$, and

$$\mathbf{h}_{i,\tau} = \gamma_\tau \sum_{\ell=0}^{L} s_\tau^{(\ell)} \mathbf{h}_i^{(\ell)}$$

Center-of-gravity:

$$E_s[\ell] = \Sigma_\ell\, \ell\, s_\tau^{(\ell)} / \Sigma_\ell\, s_\tau^{(\ell)}$$

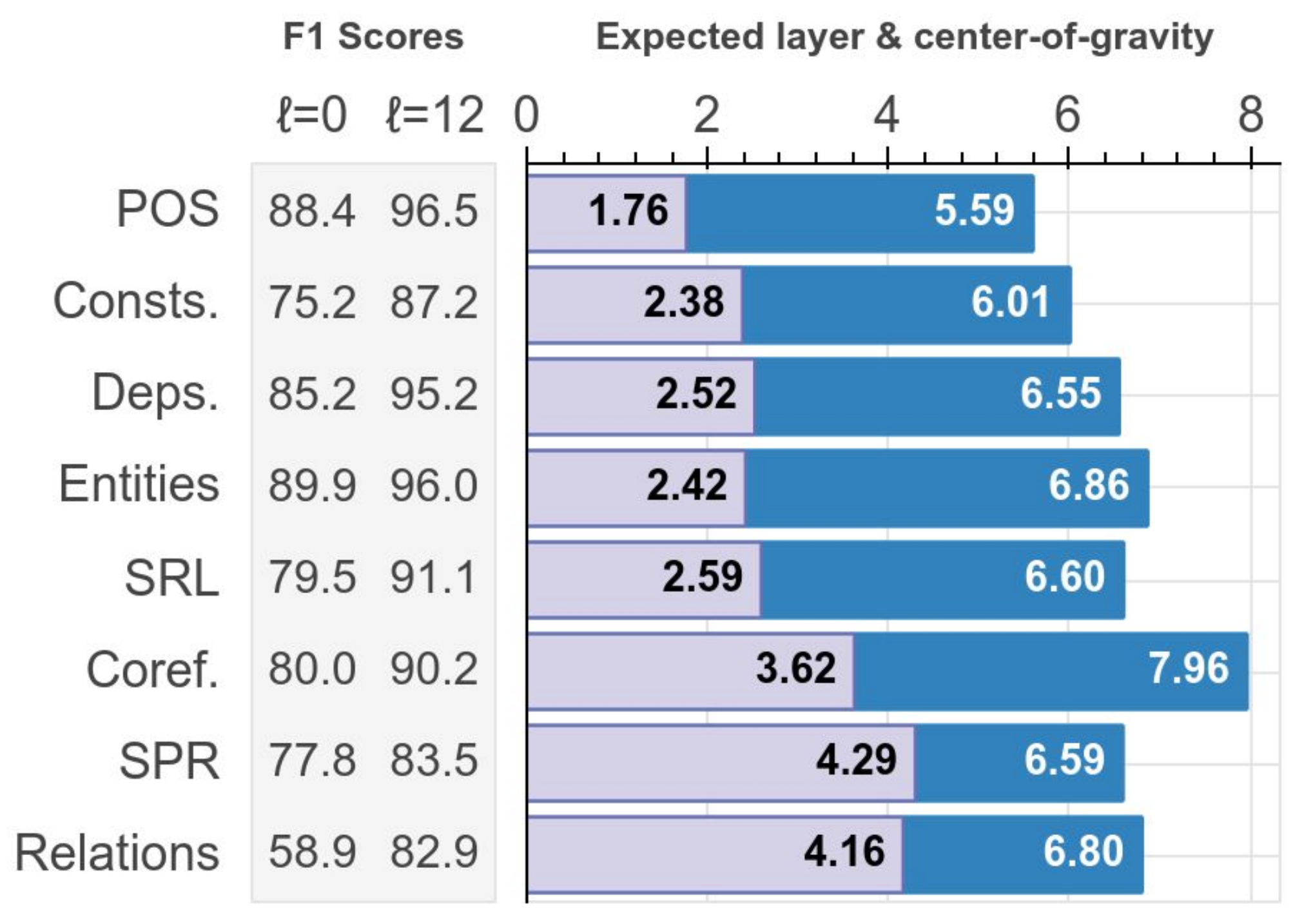
Cumulative Scoring:

$$\Delta_\tau^{(\ell)} = \mathrm{Score}(P_\tau^{(\ell)}) - \mathrm{Score}(P_\tau^{(\ell-1)})$$

Expected layer:

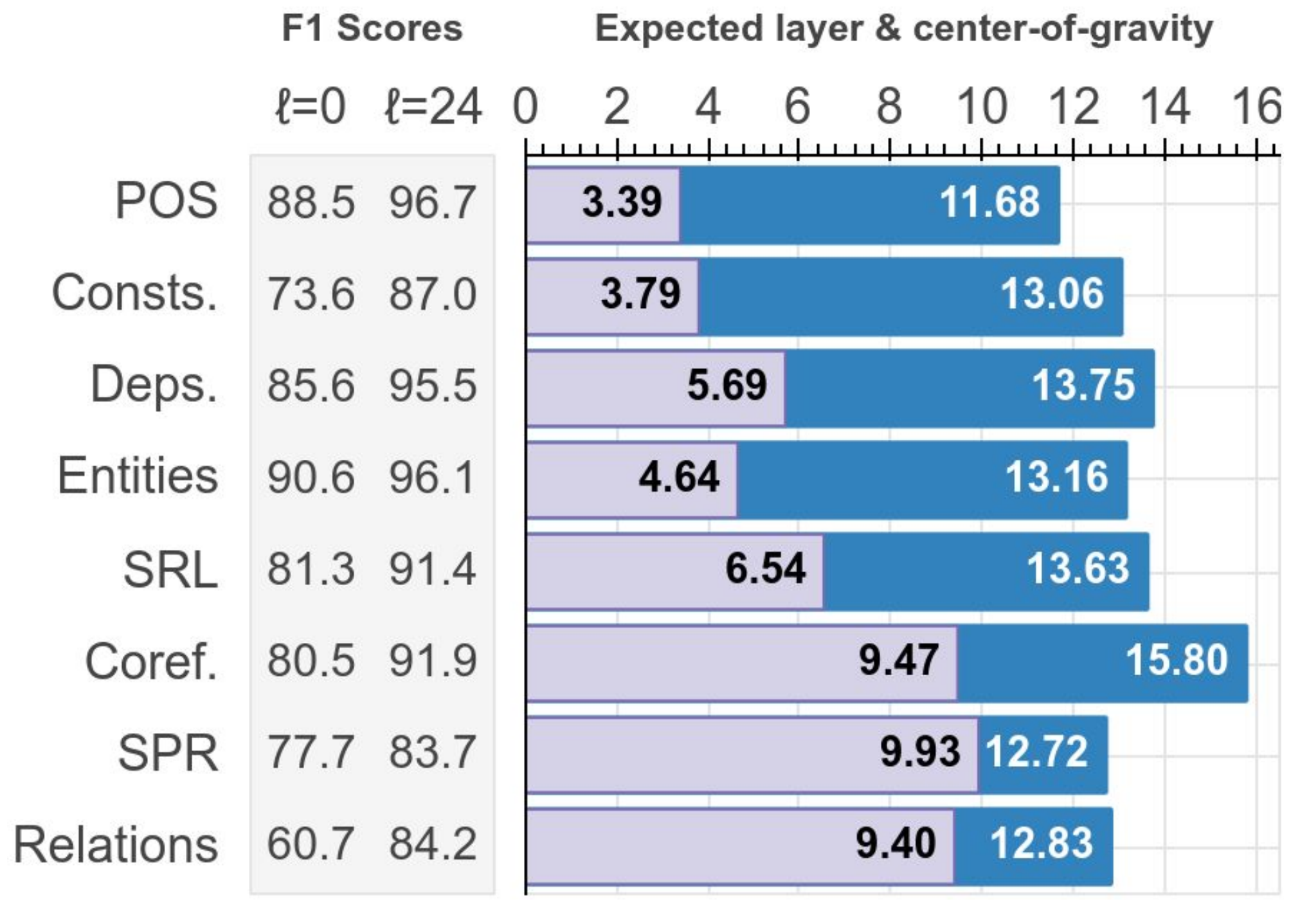
$$E_\Delta[\ell] = \Sigma_\ell\, \ell\, \Delta_\tau^{(\ell)} / \Sigma_\ell\, \Delta_\tau^{(\ell)}$$

Tasks appear in a consistent order, reflecting the traditional NLP pipeline!



	F1 Scores		Expected layer & center-of-gravity	
	ℓ=0	ℓ=12		
POS	88.4	96.5	1.76	5.59
Consts.	75.2	87.2	2.38	6.01
Deps.	85.2	95.2	2.52	6.55
Entities	89.9	96.0	2.42	6.86
SRL	79.5	91.1	2.59	6.60
Coref.	80.0	90.2	3.62	7.96
SPR	77.8	83.5	4.29	6.59
Relations	58.9	82.9	4.16	6.80

BERT-base (12 layer)



	F1 Scores		Expected layer & center-of-gravity	
	ℓ=0	ℓ=24		
POS	88.5	96.7	3.39	11.68
Consts.	73.6	87.0	3.79	13.06
Deps.	85.6	95.5	5.69	13.75
Entities	90.6	96.1	4.64	13.16
SRL	81.3	91.4	6.54	13.63
Coref.	80.5	91.9	9.47	15.80
SPR	77.7	83.7	9.93	12.72
Relations	60.7	84.2	9.40	12.83

BERT-large (24 layer)

Tracing a Sentence

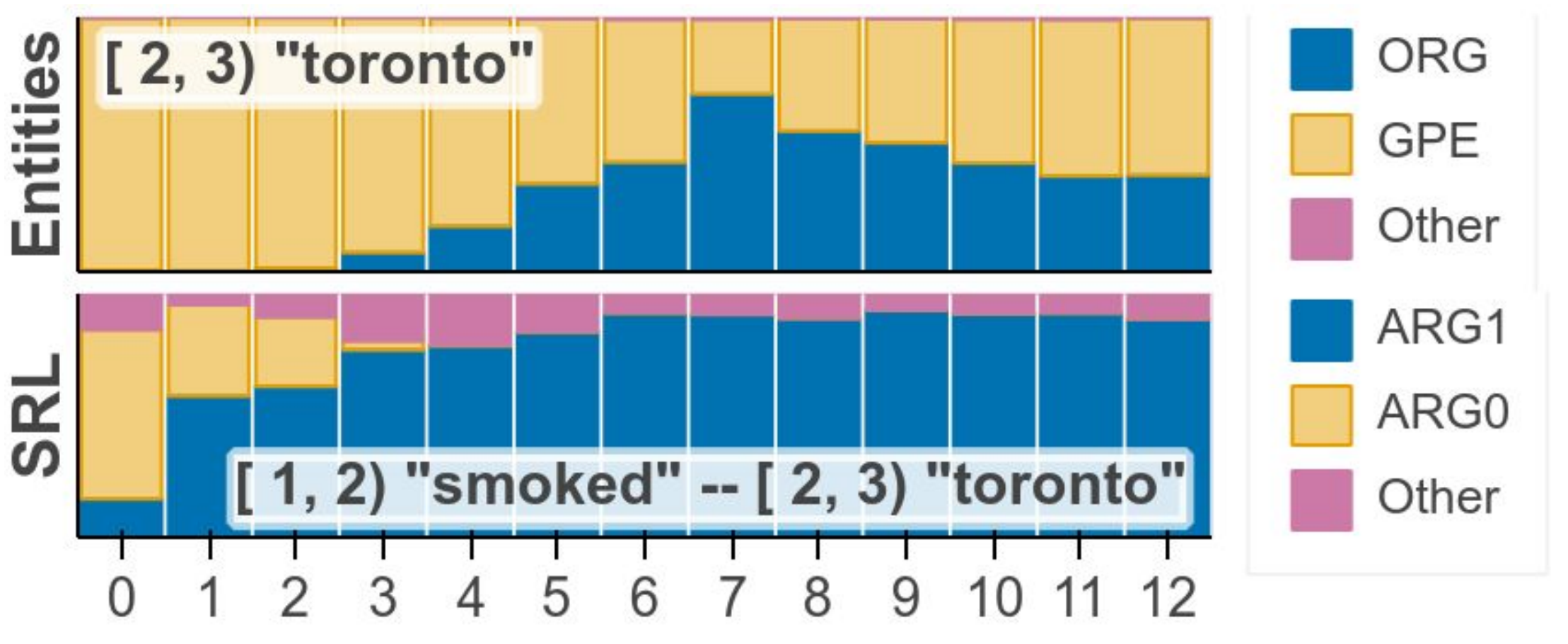
OntoNotes: τ = {POS, constituents, entities, SRL, coref}

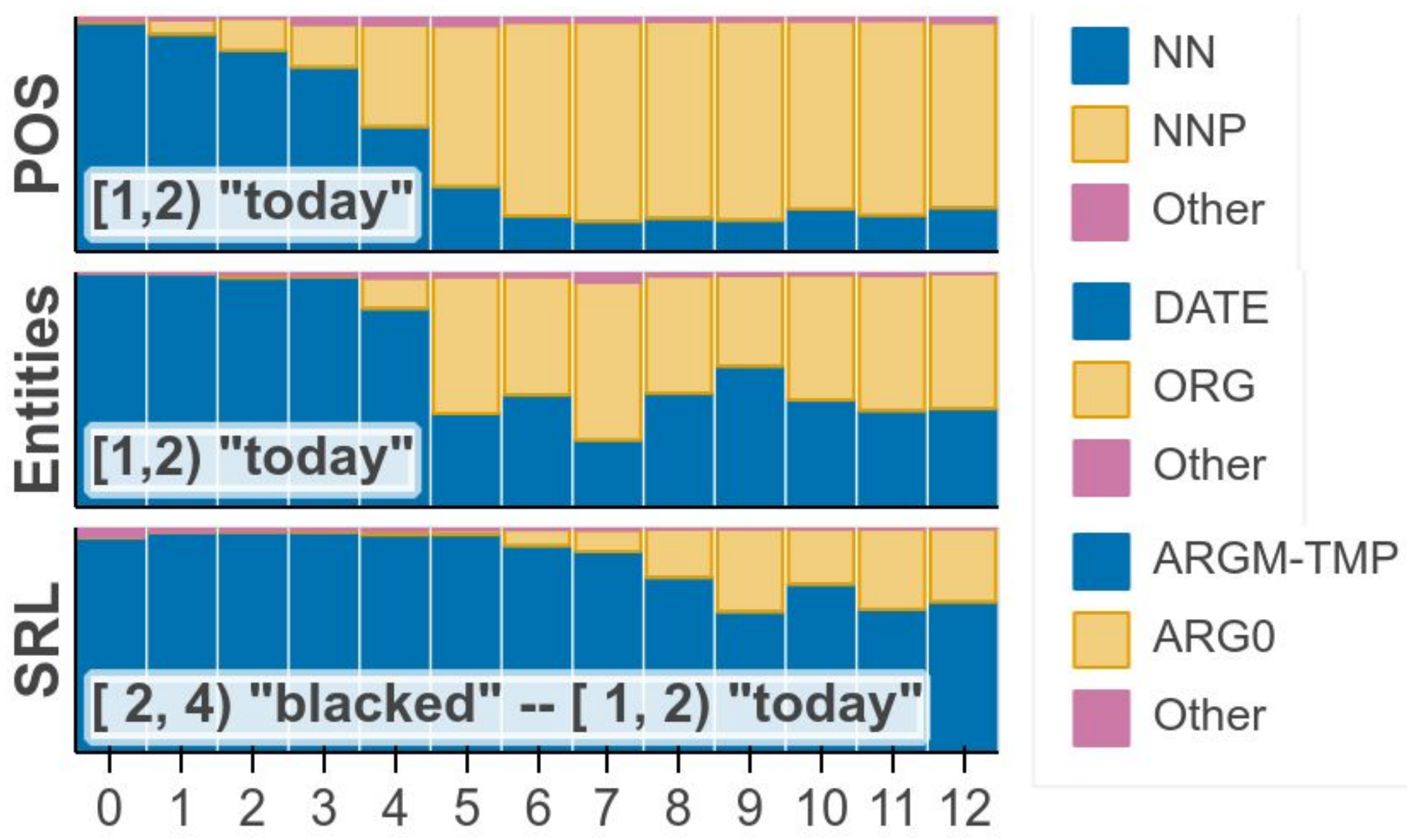
Collect predictions $\{P_\tau^{(\ell)}\}$ for $\ell = 0, 1, \ldots, L$ for each task

“he smoked **toronto** in the playoffs with six hits, … ”

Entities: from **GPE** → **ORG** in layers 3-7

SRL: conclude **ARG1** by layer 2-3





“china **today** blacked out a cnn interview that was critical … ”

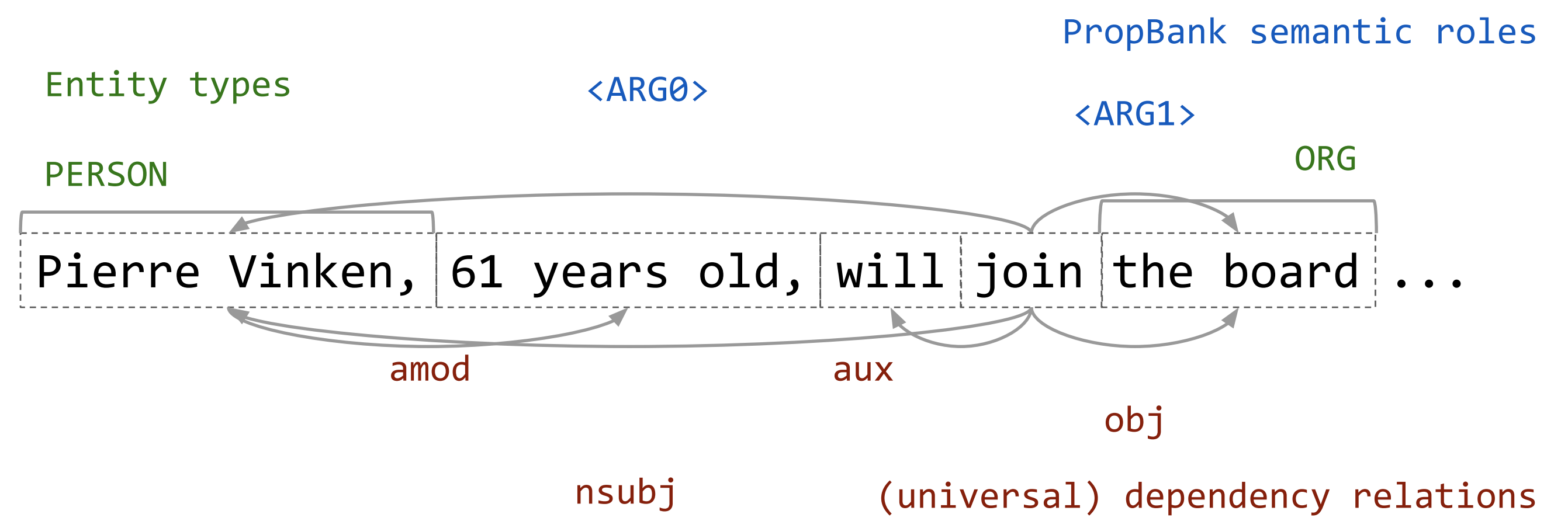
POS: from **NN** → **NNP** in layers 3-5

Entities: from **DATE** → **ORG** in layer 4-5

SRL: consider **ARG0** from layers 6-9

Edge Probing

Probing suite [2] recasts tasks as edge labeling:



PropBank semantic roles

Entity types

<ARG0>

<ARG1>

PERSON

ORG

Pierre Vinken, 61 years old, will join the board ...

amod

aux

obj

nsubj

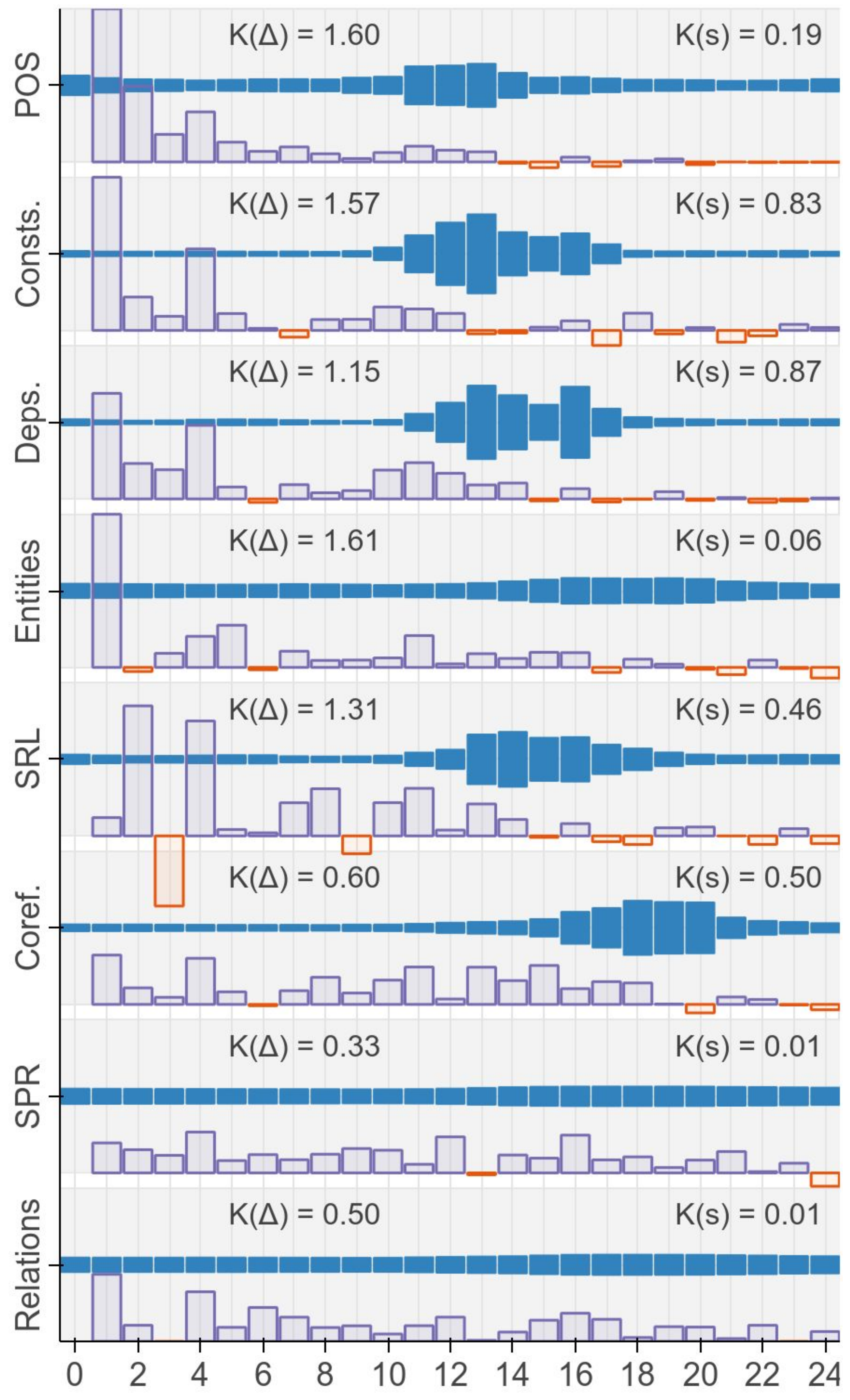
(universal) dependency relations

Given contextual vectors $E = [e_0, e_1, \ldots, e_n]$, predict:

- **Unary:** label(s) for span1 = $[i_1, j_1)$
- **Binary:** label(s) for (span1 = $[i_1, j_1)$, span2 = $[i_2, j_2)$)

Common classifier model [2] over frozen encoder, with ELMo-style mixing over layers {0,1,...,ℓ}.

Per-layer Contributions



Solid blue: scalar mixing weights (s)

Light purple: relative improvement in F1 score (Δ)

K(*): KL(* | Uniform) over all layers

References

[1] *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (Devlin et al., NAACL 2019)

[2] *What do you learn from context? Probing for sentence structure in contextualized word representations* (Tenney et al., ICLR 2019)