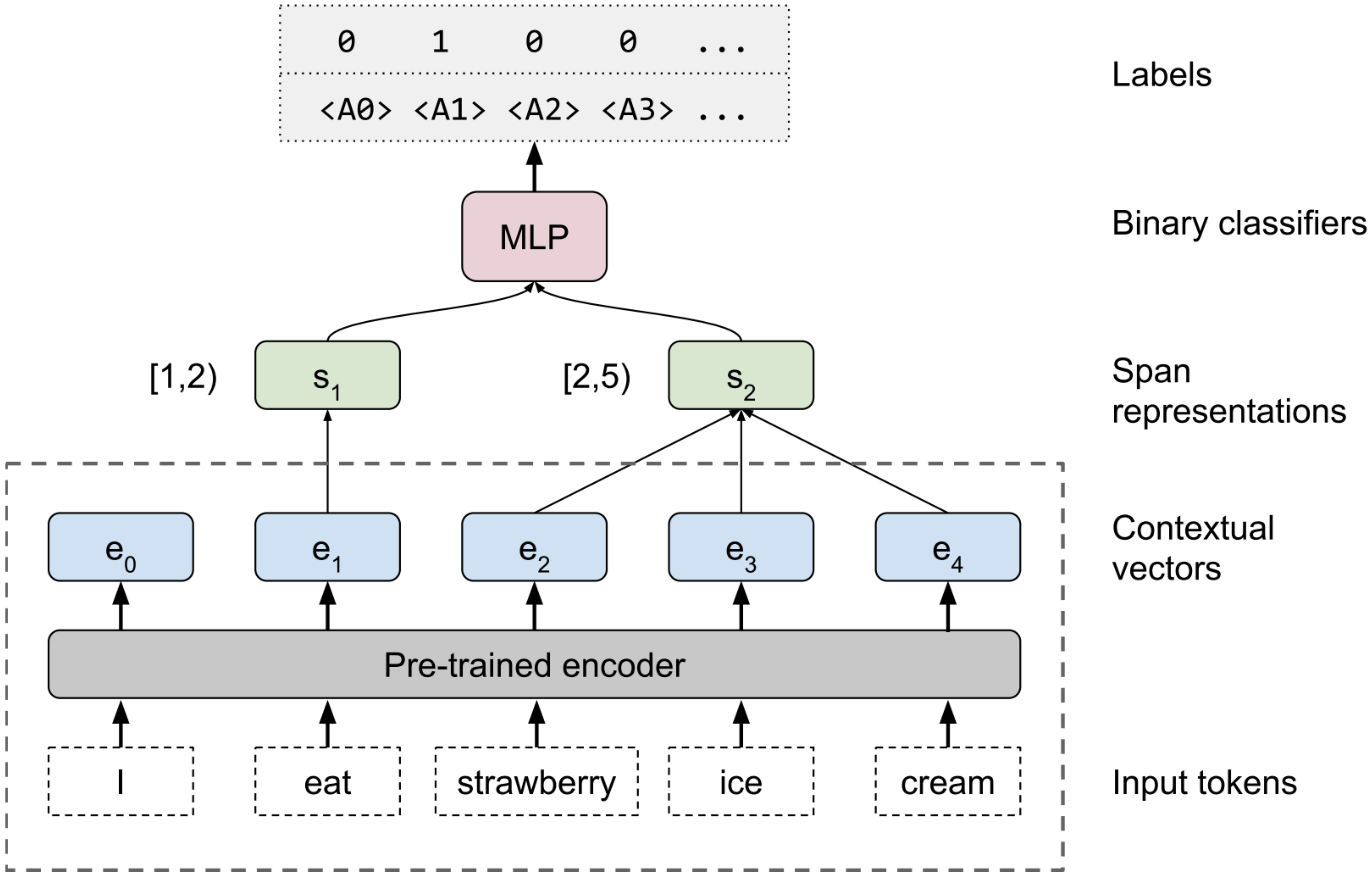
Probing

- We want to know what information is captured in a neural network. Try to predict that information from the network's representations

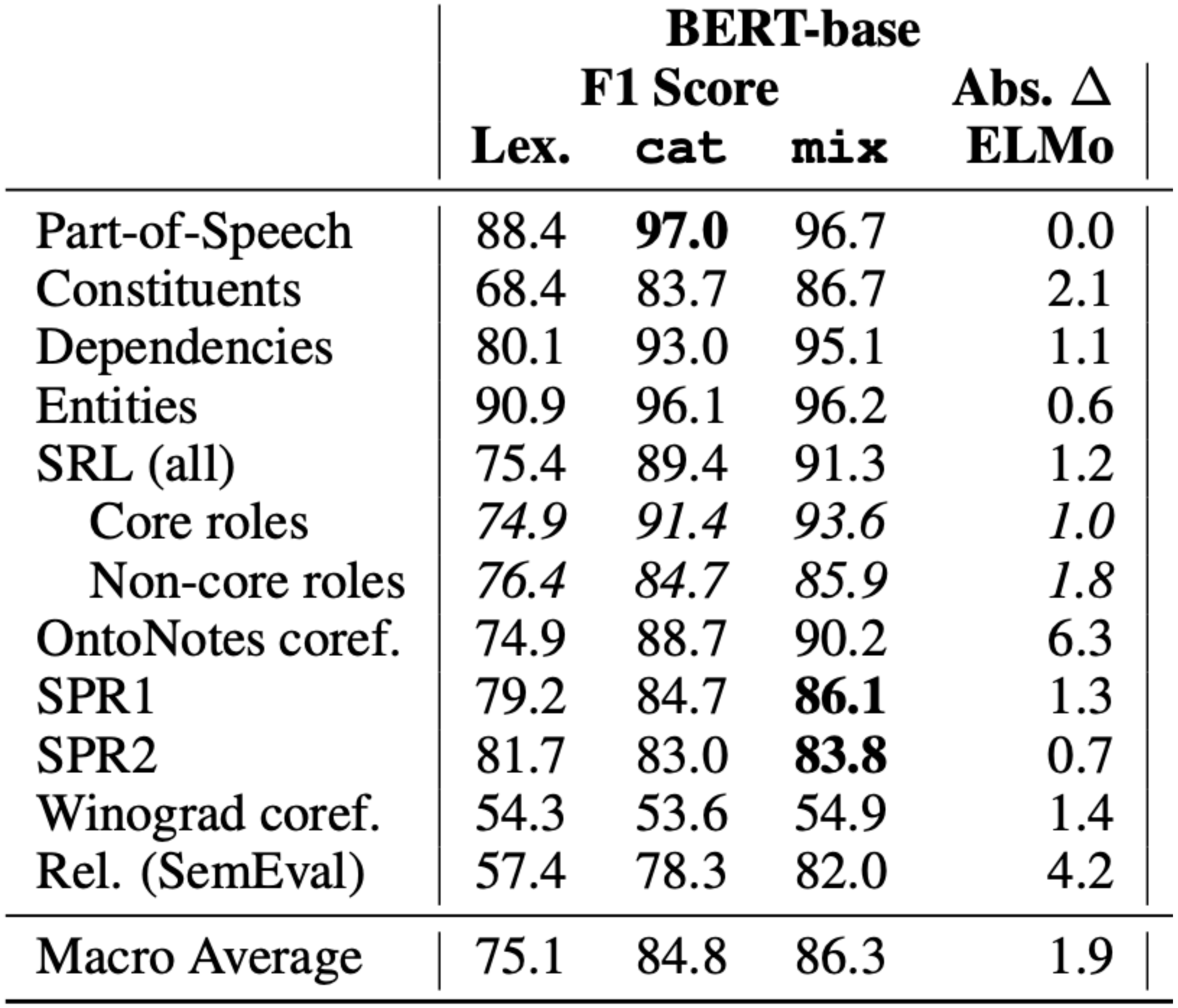


- Given a simple, fixed class of model (e.g., one-layer FFNN), how well can we predict various things from word representations?

Tenney et al. (2019)

Probing: Results

- Lex: baseline built on context-independent vectors
- Large gains from contextualization, and BERT beats ELMo

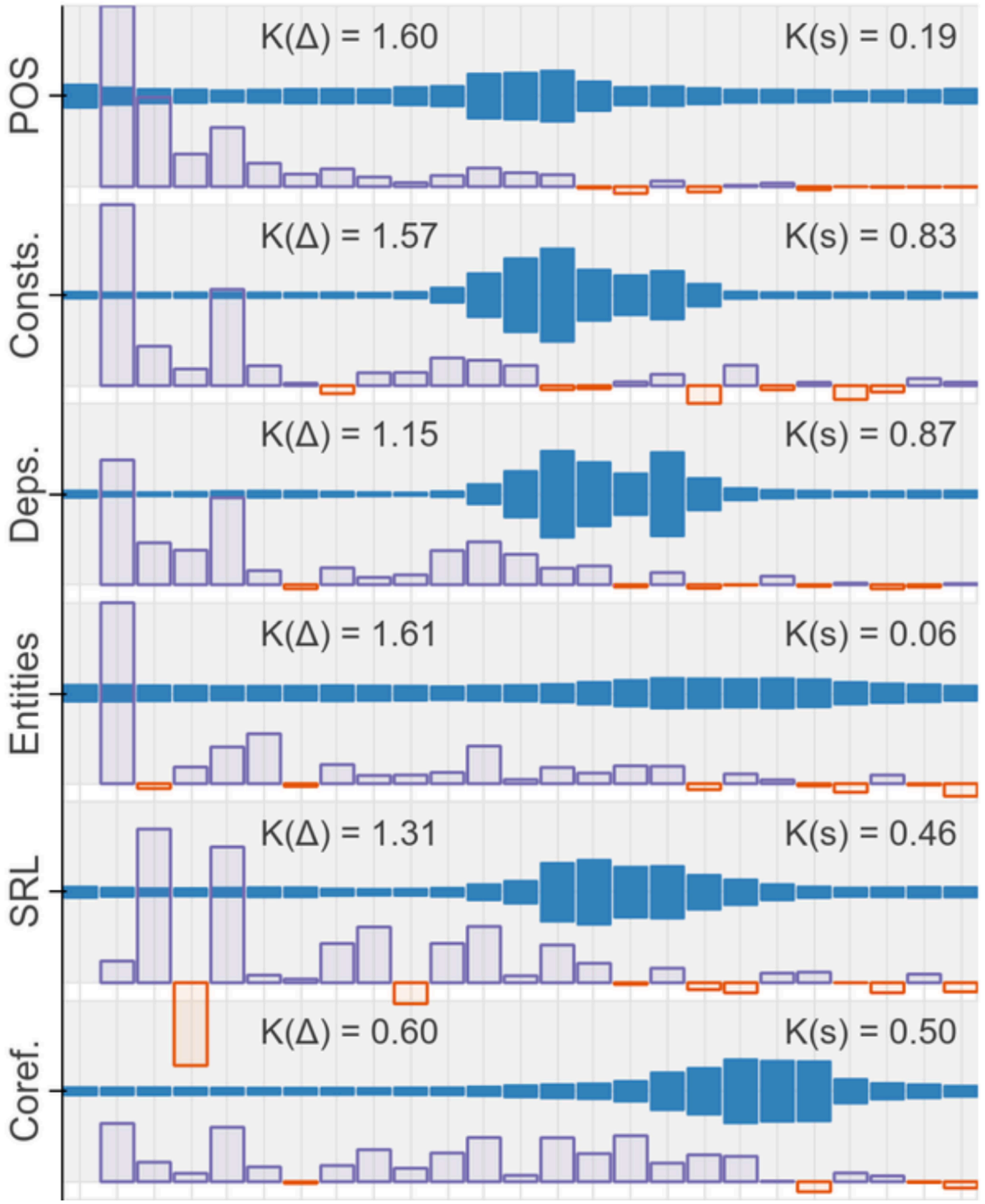


	BERT-base			
	F1 Score			Abs. Δ
	Lex.	cat	mix	ELMo
Part-of-Speech	88.4	**97.0**	96.7	0.0
Constituents	68.4	83.7	86.7	2.1
Dependencies	80.1	93.0	95.1	1.1
Entities	90.9	96.1	96.2	0.6
SRL (all)	75.4	89.4	91.3	1.2
Core roles	*74.9*	*91.4*	*93.6*	*1.0*
Non-core roles	*76.4*	*84.7*	*85.9*	*1.8*
OntoNotes coref.	74.9	88.7	90.2	6.3
SPR1	79.2	84.7	**86.1**	1.3
SPR2	81.7	83.0	**83.8**	0.7
Winograd coref.	54.3	53.6	54.9	1.4
Rel. (SemEval)	57.4	78.3	82.0	4.2
Macro Average	75.1	84.8	86.3	1.9

Tenney et al. (2019)

Probing: Results

- Purple: BERT-large performance on each task (as delta from mean) using representations from that layer of BERT
- Earlier layers of the network: better at POS and low-level tasks. Later layers are better at higher-level tasks like coreference



Tenney et al. (2019b)