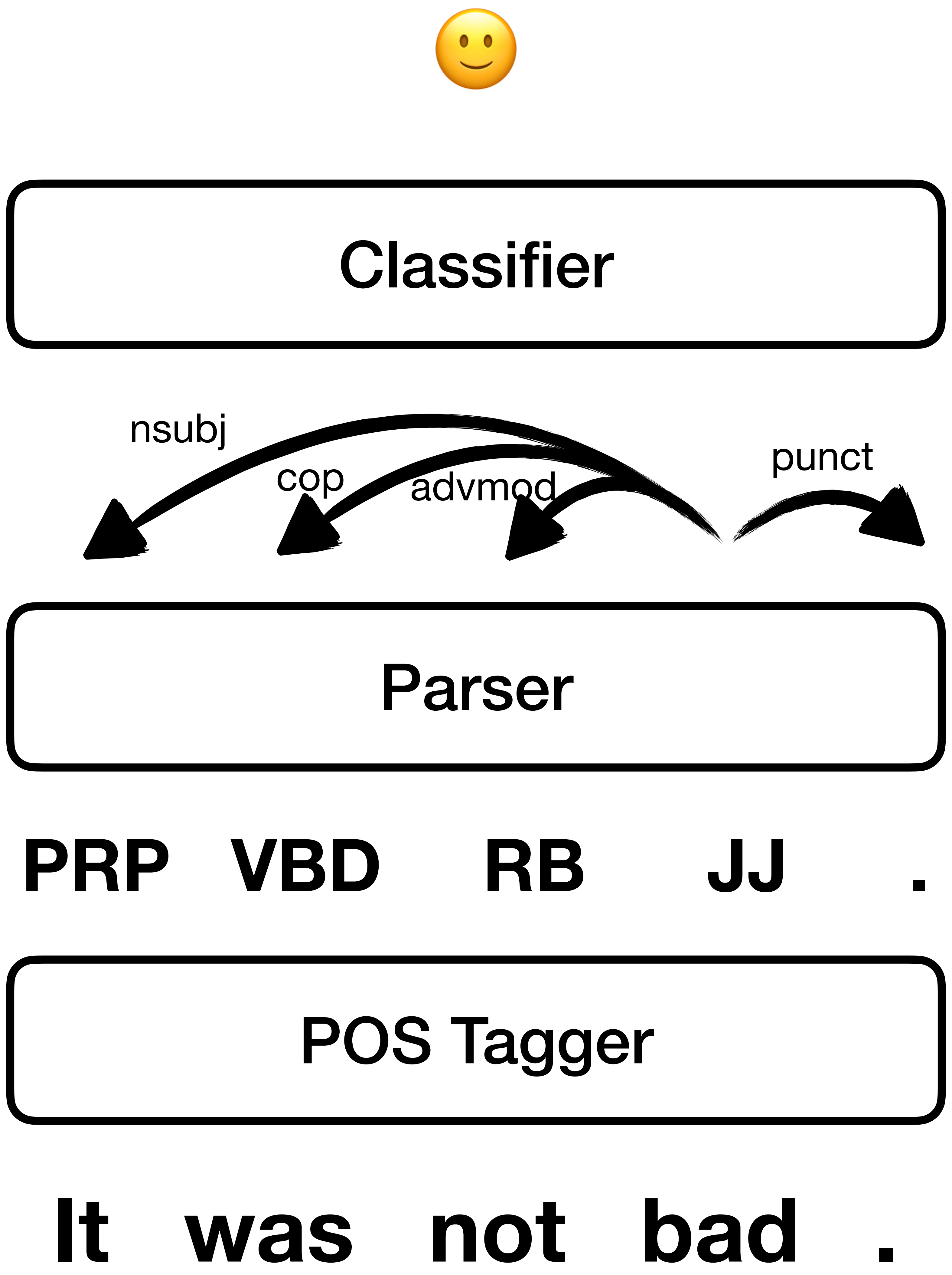
Infusing Finetuning with Semantic Dependencies

Zhaofeng Wu, Hao Peng, and Noah Smith

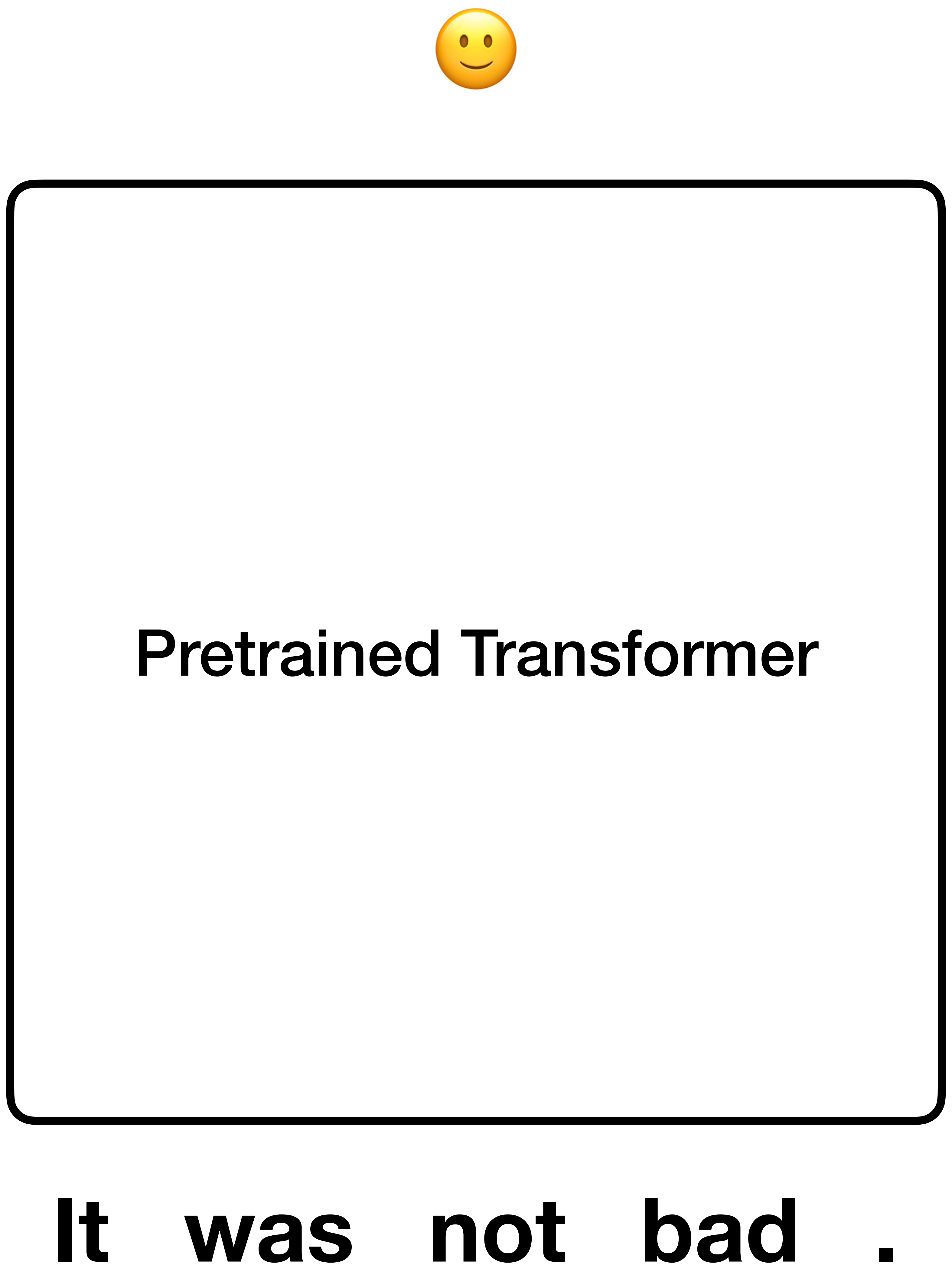
TACL 2021; 9:226–242

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

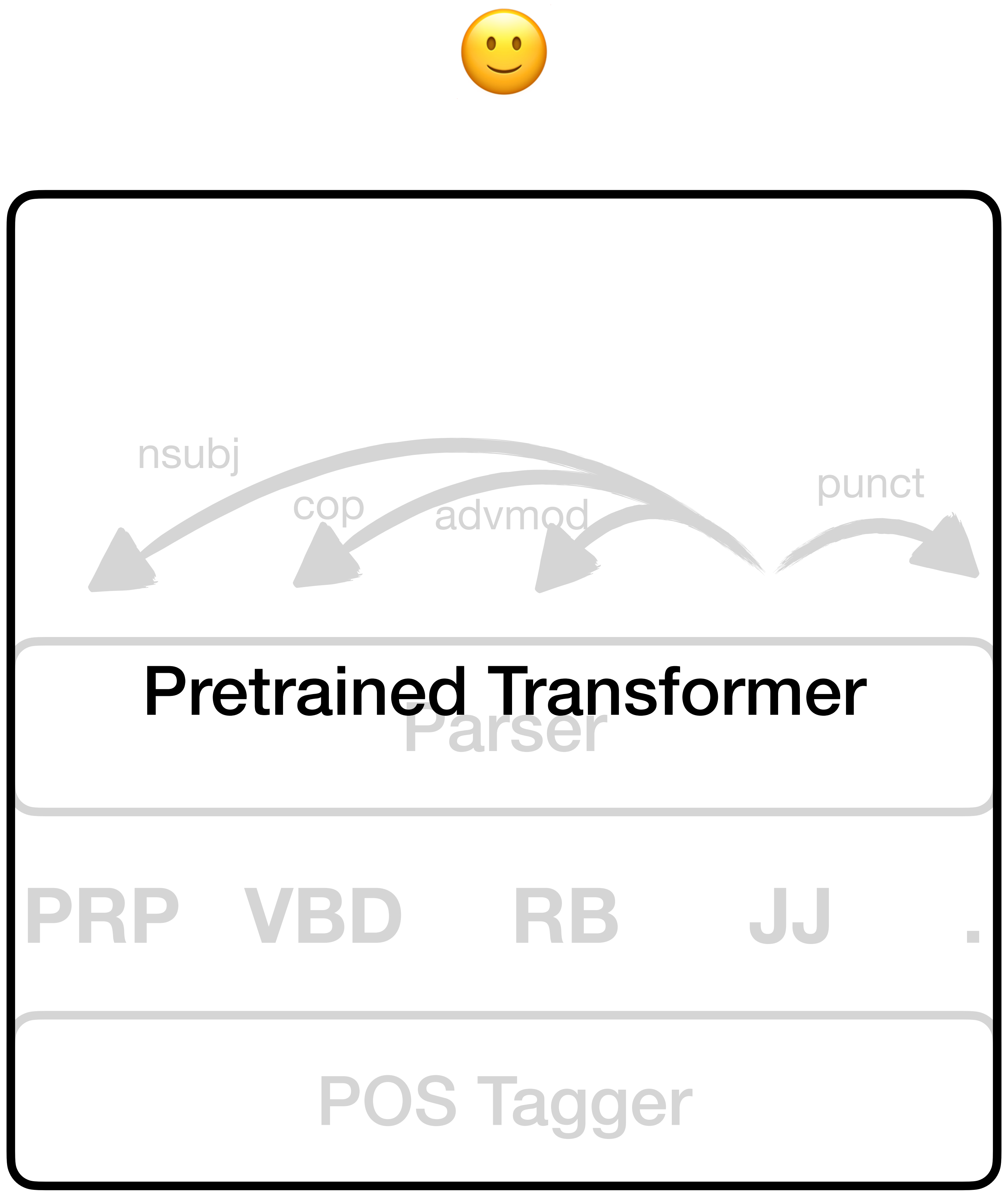
Motivation


Classifier
nsubj
cop
advmod
punct
Parser
PRP VBD RB JJ .
POS Tagger
It was not bad .

Motivation



Motivation



It was not bad .

Would semantics help?

Introduction

Introduction

- We show BERT/RoBERTa less prominently surface semantics…

Introduction

- We show BERT/RoBERTa less prominently surface semantics…
- … and the explicit incorporation of semantic information:

Introduction

- We show BERT/RoBERTa less prominently surface semantics…
- … and the explicit incorporation of semantic information:

  1. Improves downstream task performance

Introduction

- We show BERT/RoBERTa less prominently surface semantics…
- … and the explicit incorporation of semantic information:

  1. Improves downstream task performance

  2. Helps guard against frequent yet invalid heuristics

Introduction

- We show BERT/RoBERTa less prominently surface semantics…
- … and the explicit incorporation of semantic information:
  1. Improves downstream task performance
  2. Helps guard against frequent yet invalid heuristics
  3. Better captures nuanced linguistic phenomena

Introduction

- We show BERT/RoBERTa less prominently surface semantics…
- … and the explicit incorporation of semantic information:

1. Improves downstream task performance
2. Helps guard against frequent yet invalid heuristics
3. Better captures nuanced linguistic phenomena
4. Increases training sample efficiency

Operationalizing “Meaning”

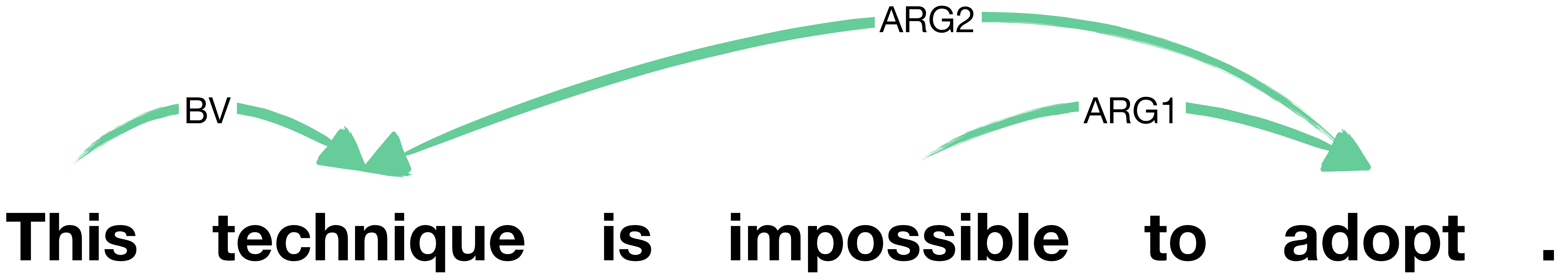
ARG2

BV ARG1

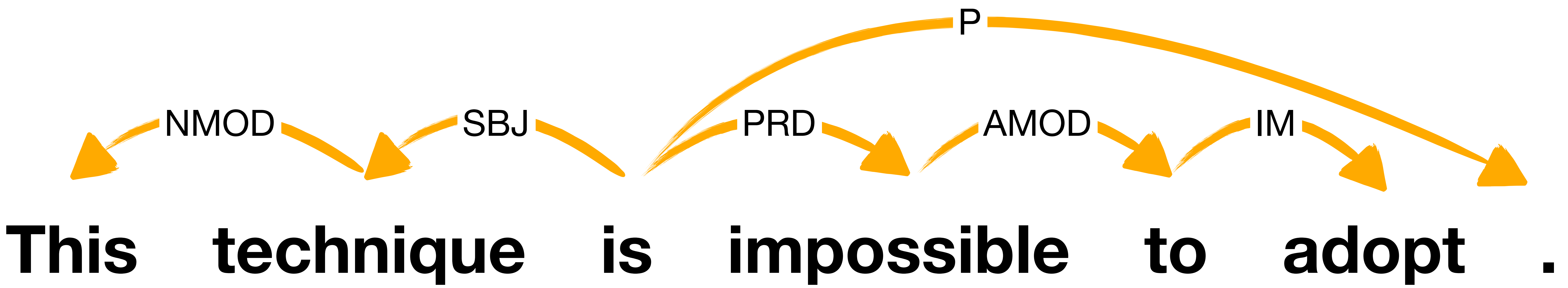
This technique is impossible to adopt .

DELPH-IN MRS-Derived Dependencies (**DM**; Ivanova et al., 2012)

Operationalizing “Meaning”

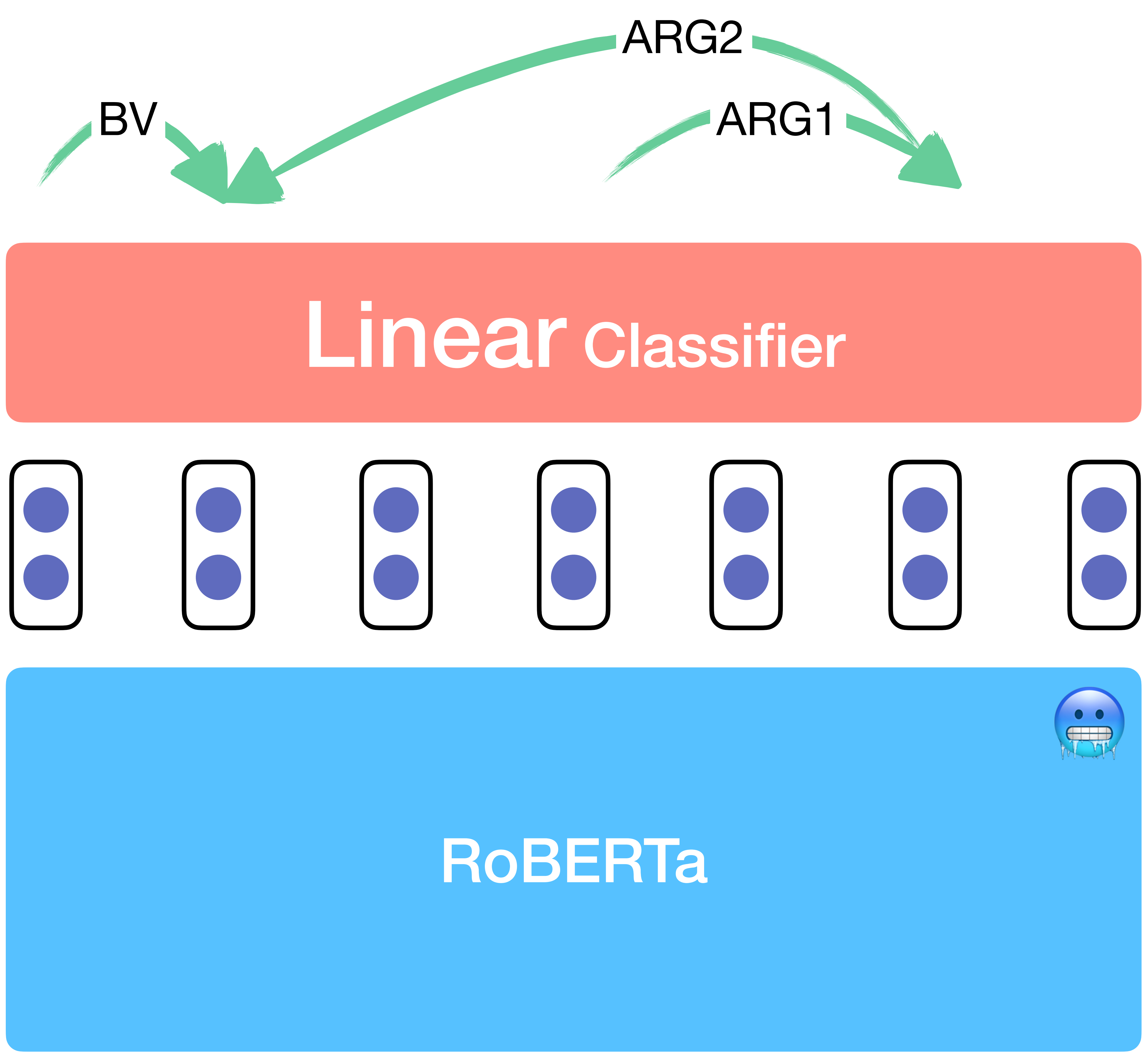


DELPH-IN MRS-Derived Dependencies (**DM**; Ivanova et al., 2012)



Stanford Dependencies (**SD**; de Marneffe et al., 2006)

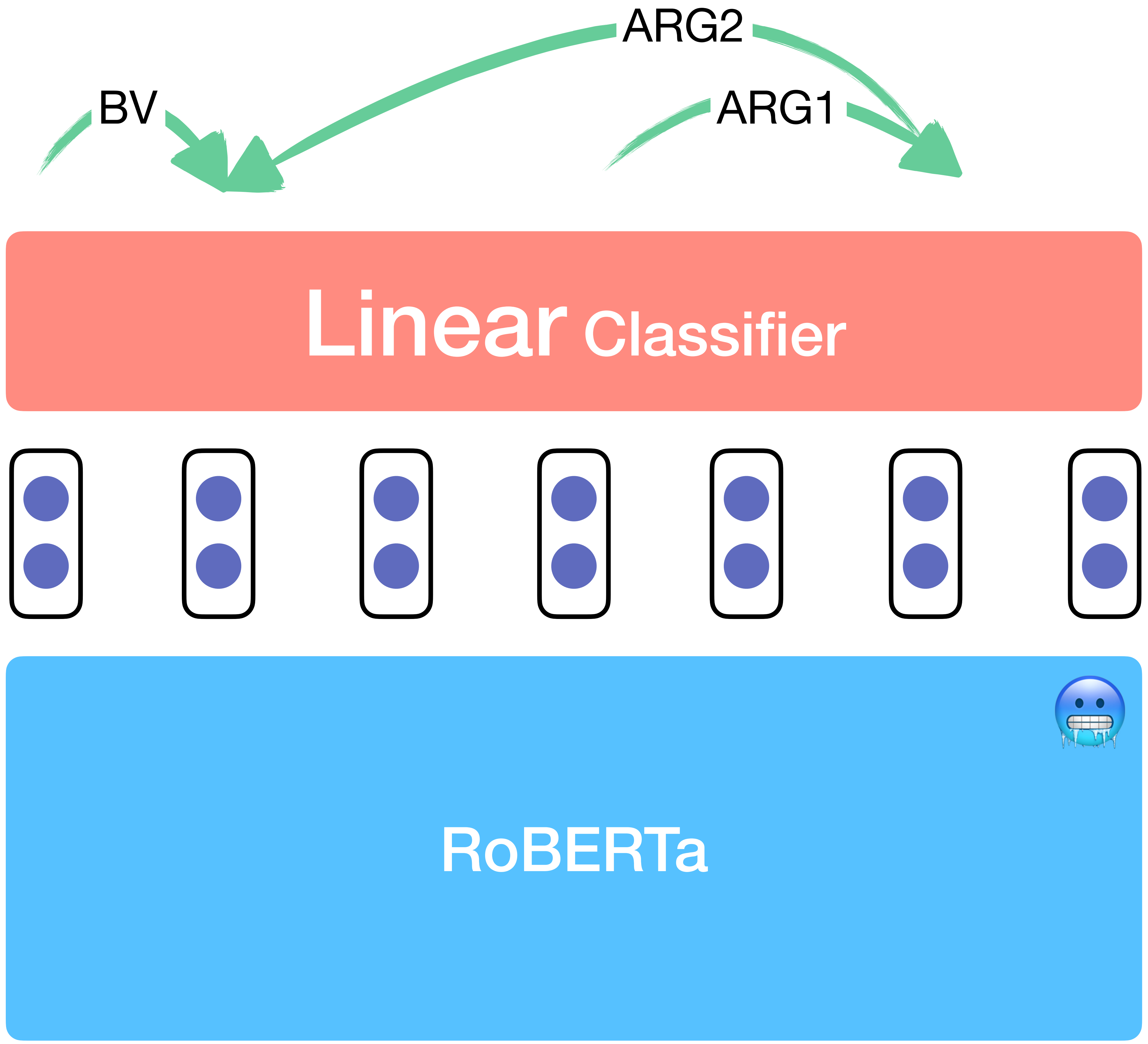
Probing RoBERTa with Semantics



This technique is impossible to adopt .

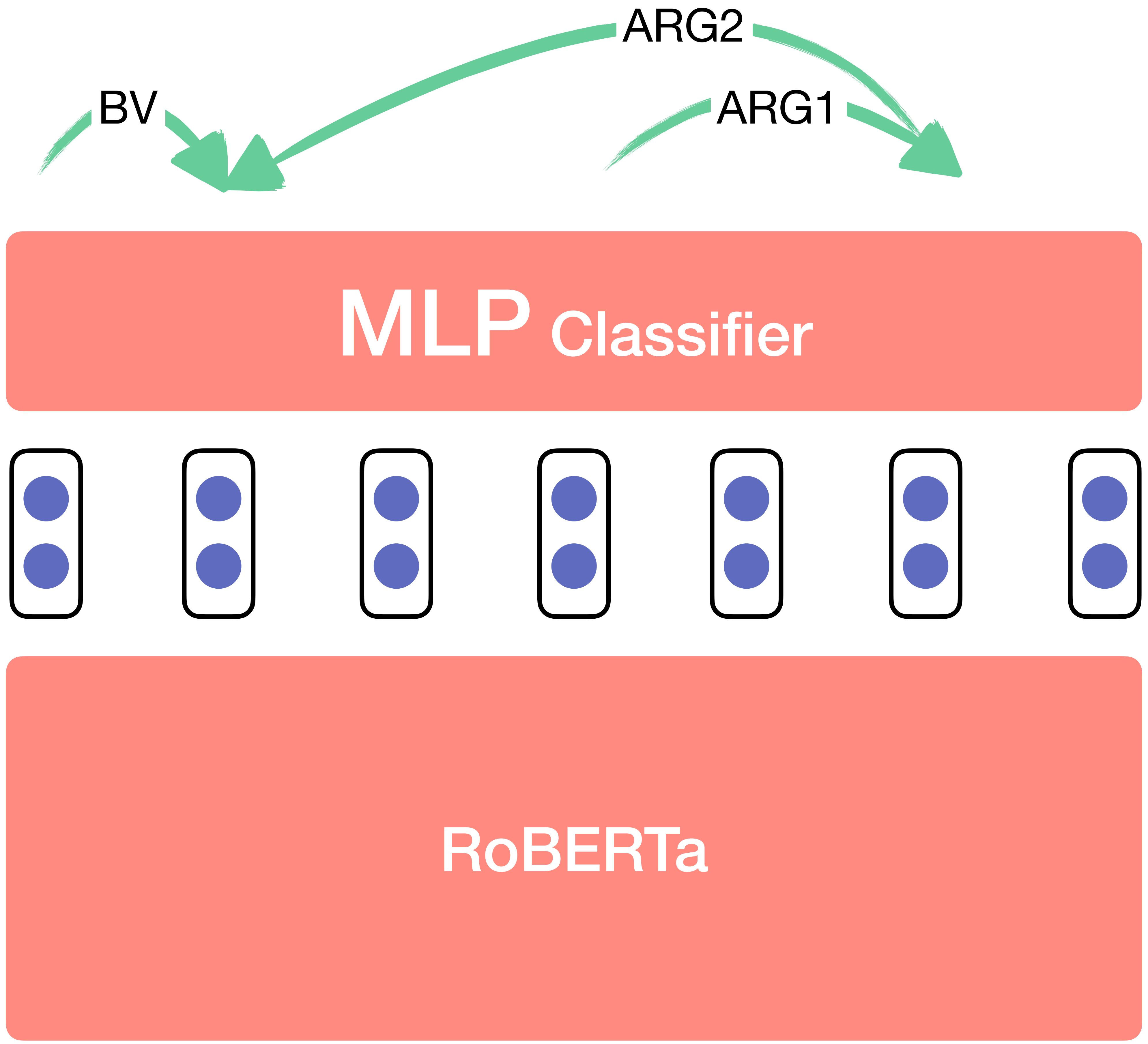
Probing model (Shi et al., 2016; Adi et al., 2017)

Probing RoBERTa with Semantics



This technique is impossible to adopt .

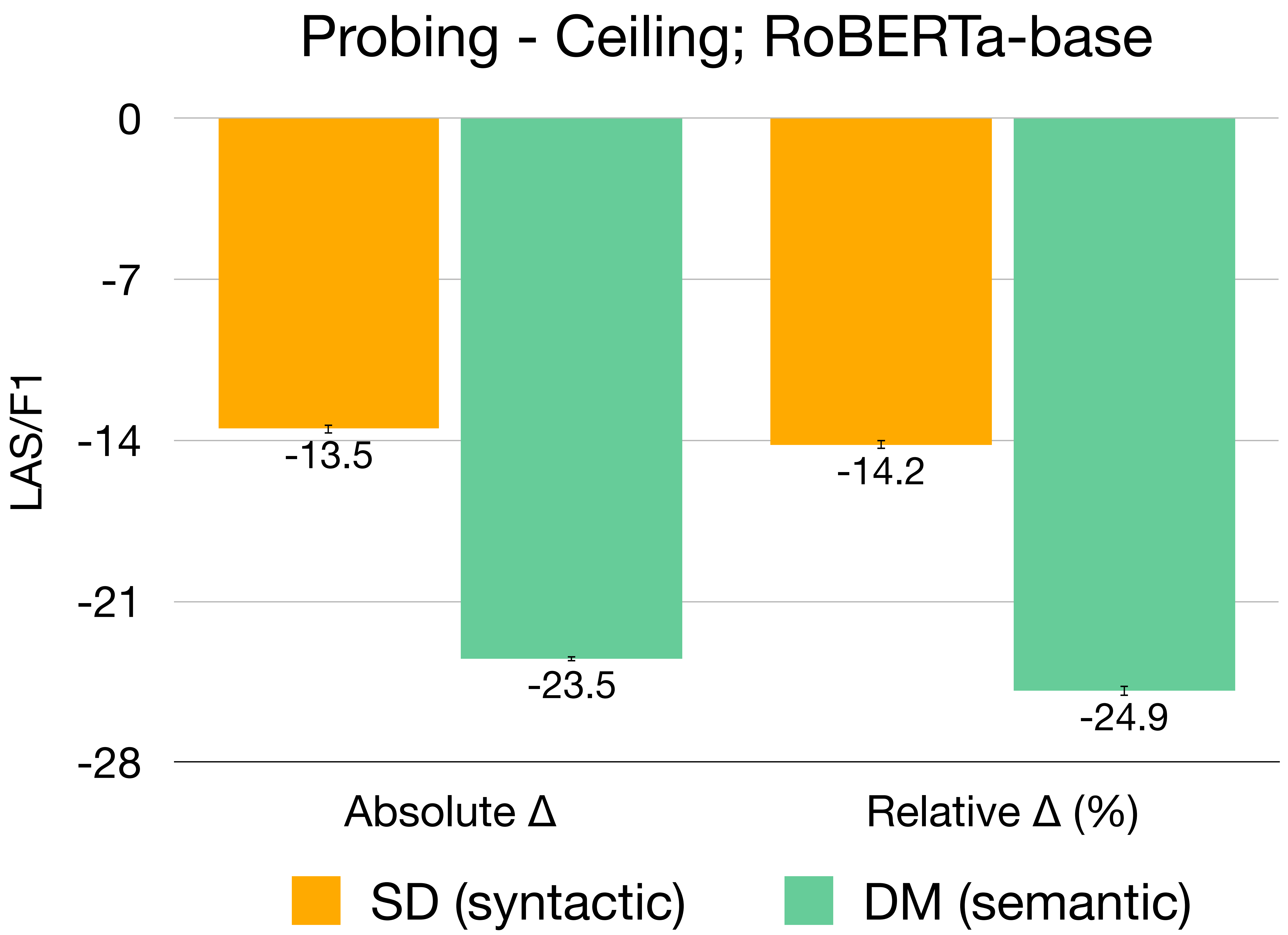
Probing model (Shi et al., 2016; Adi et al., 2017)



This technique is impossible to adopt .

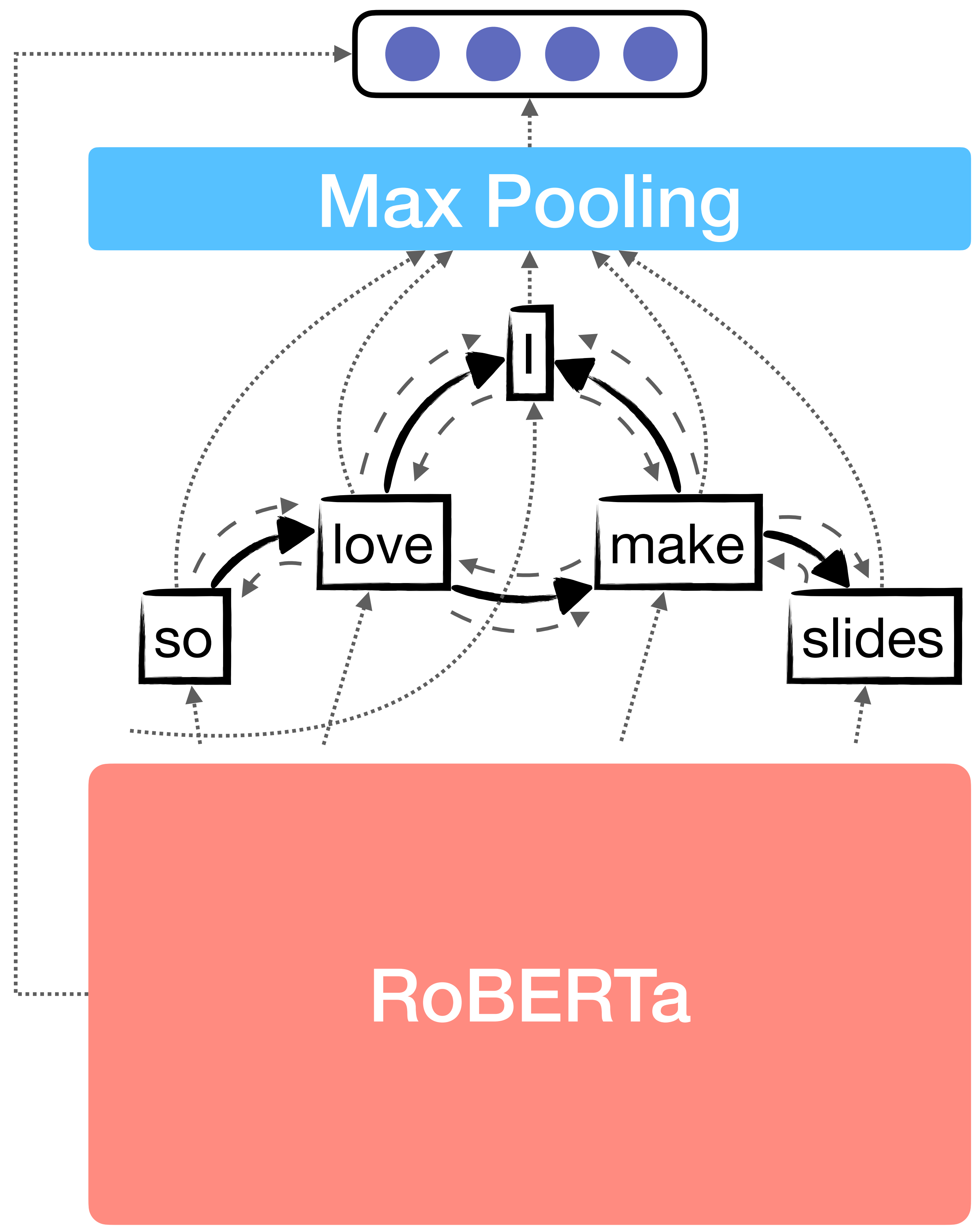
Ceiling model (Dozat and Manning, 2017, 2018)

Probing RoBERTa with Semantics



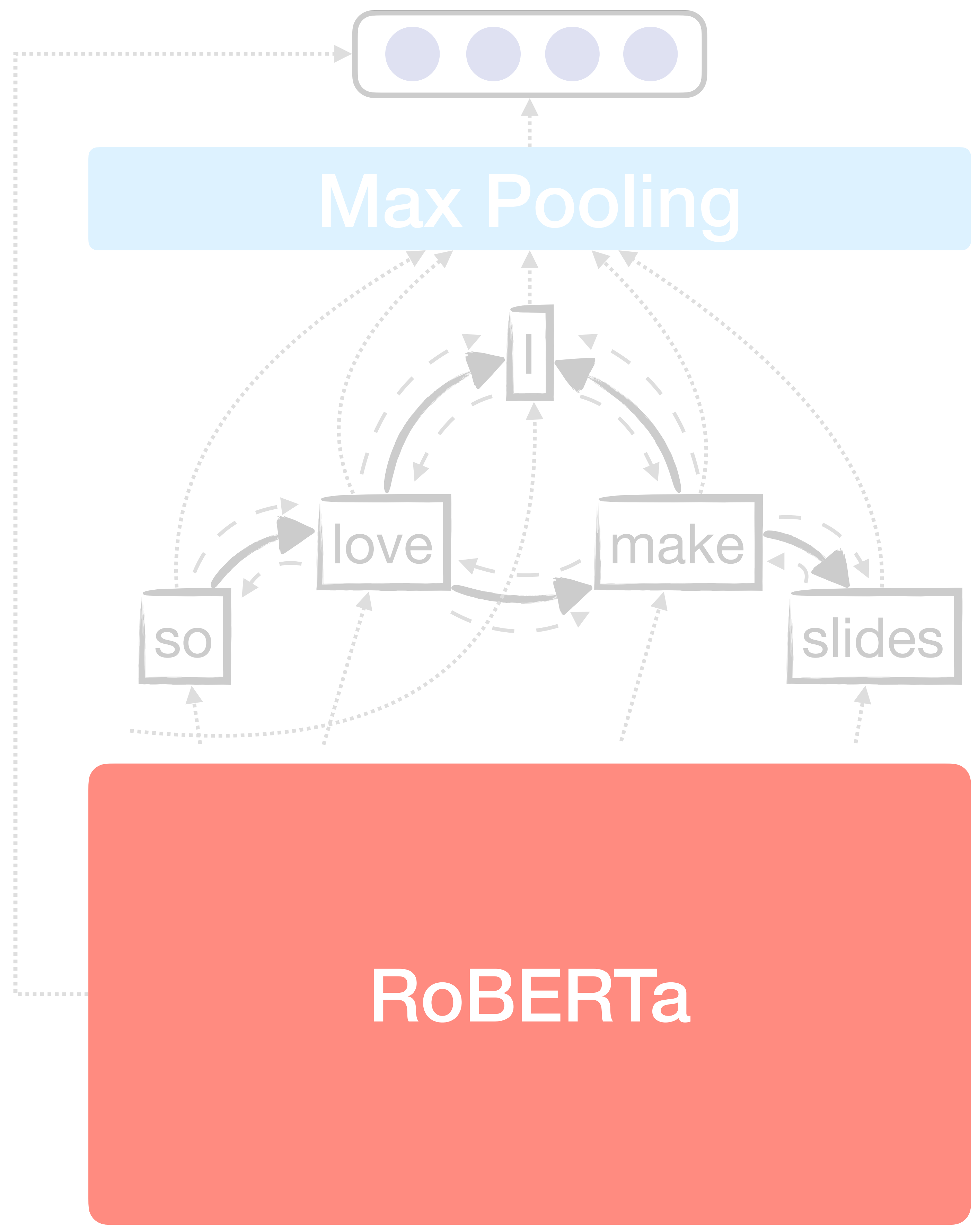
Can we use semantics to augment pretrained transformers?

Semantics-Infused Finetuning (SIFT)



I so love to make slides .

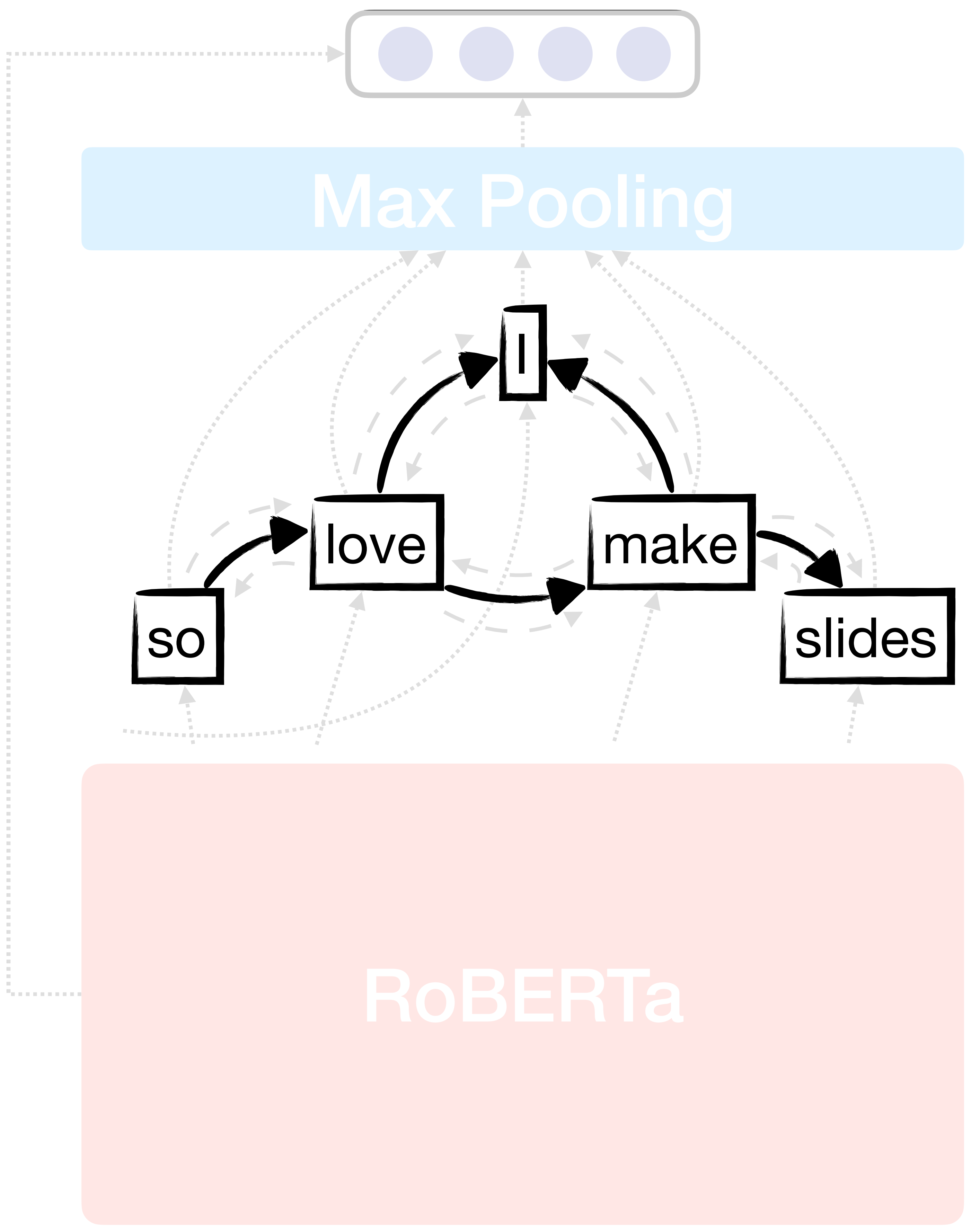
Semantics-Infused Finetuning (SIFT)



I so love to make slides .

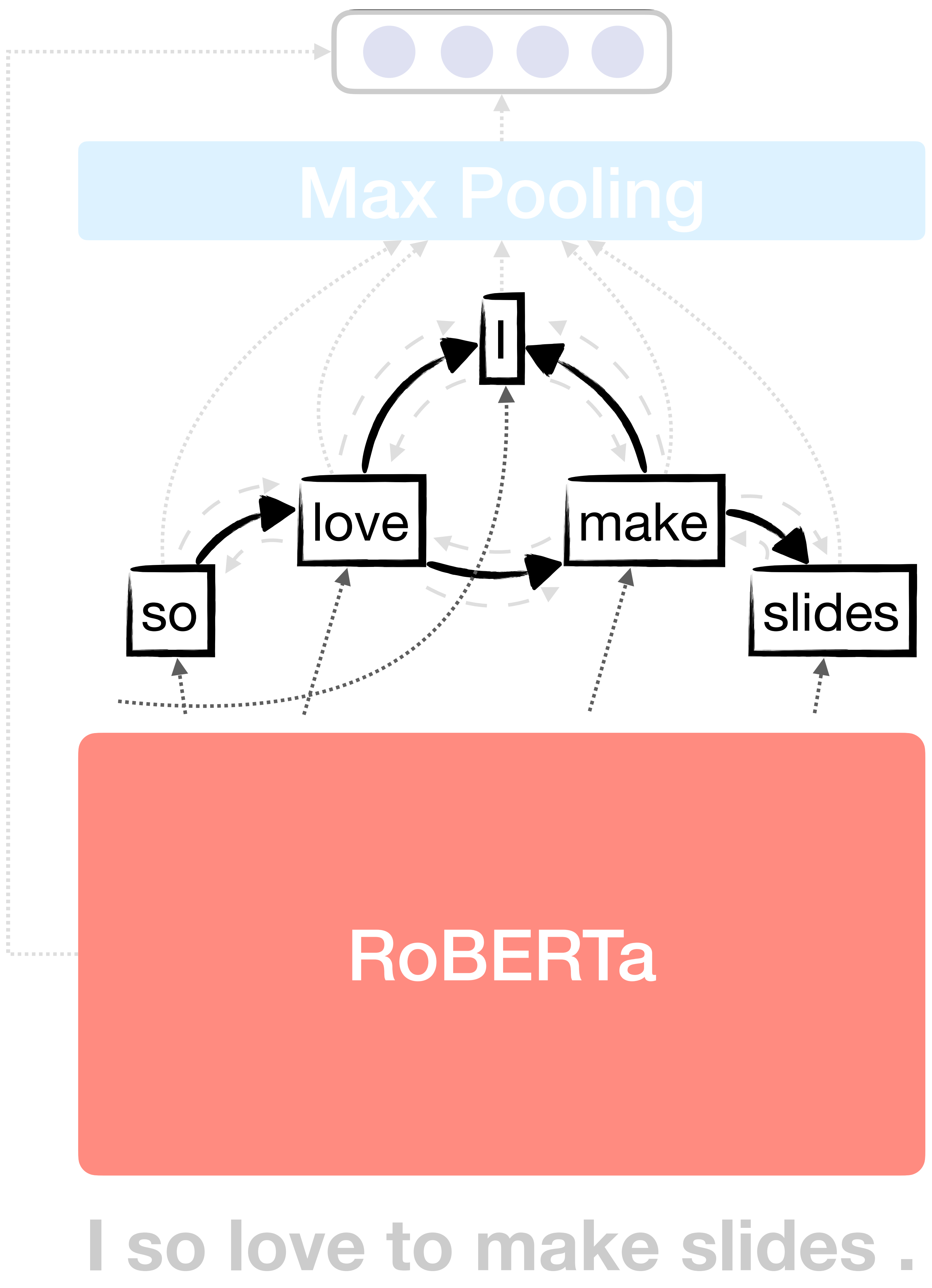
Che et al. (2019)

Semantics-Infused Finetuning (SIFT)



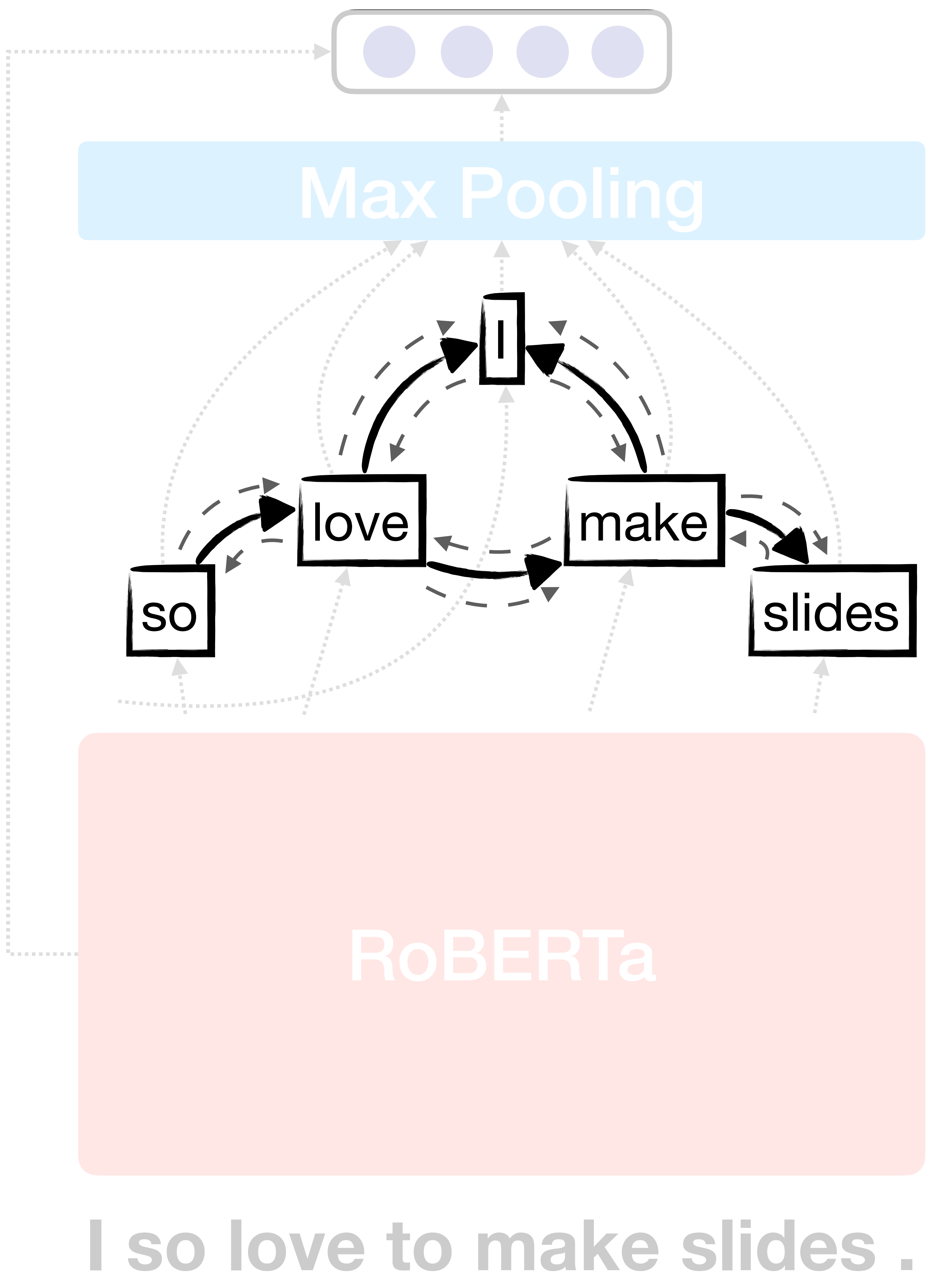
I so love to make slides .

Semantics-Infused Finetuning (SIFT)

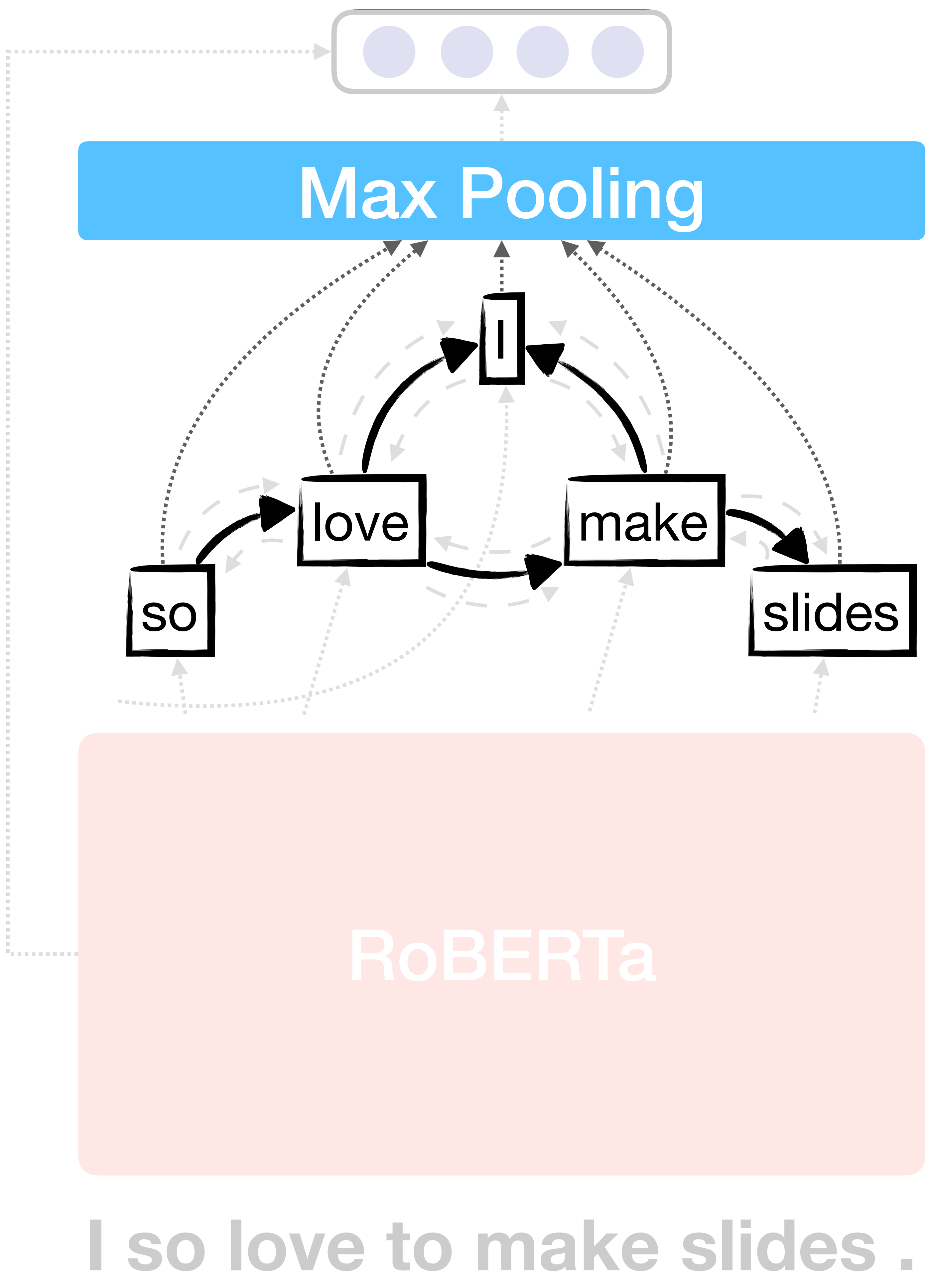


Schlichtkrull et al. (2017)

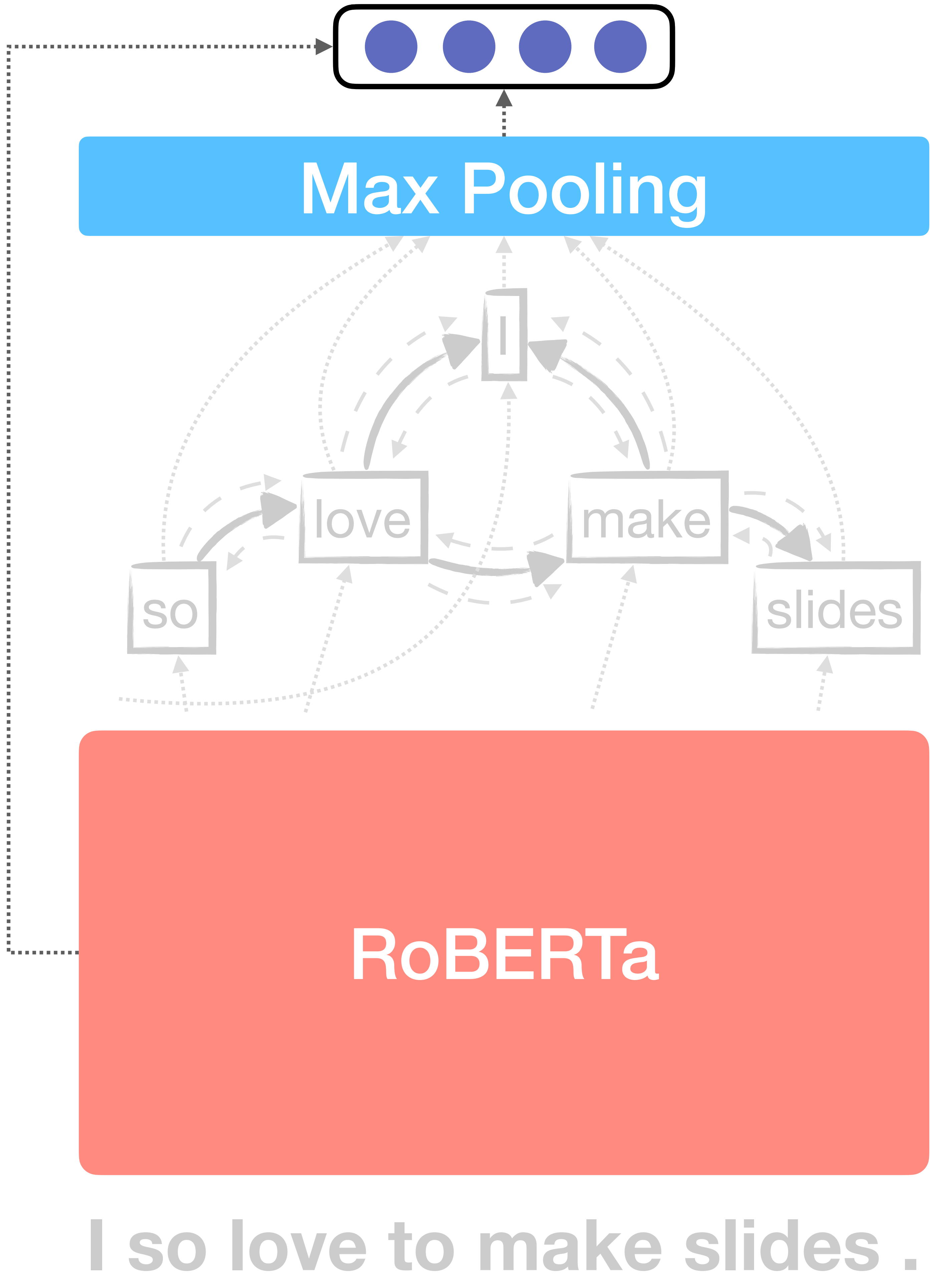
Semantics-Infused Finetuning (SIFT)



Semantics-Infused Finetuning (SIFT)



Semantics-Infused Finetuning (SIFT)

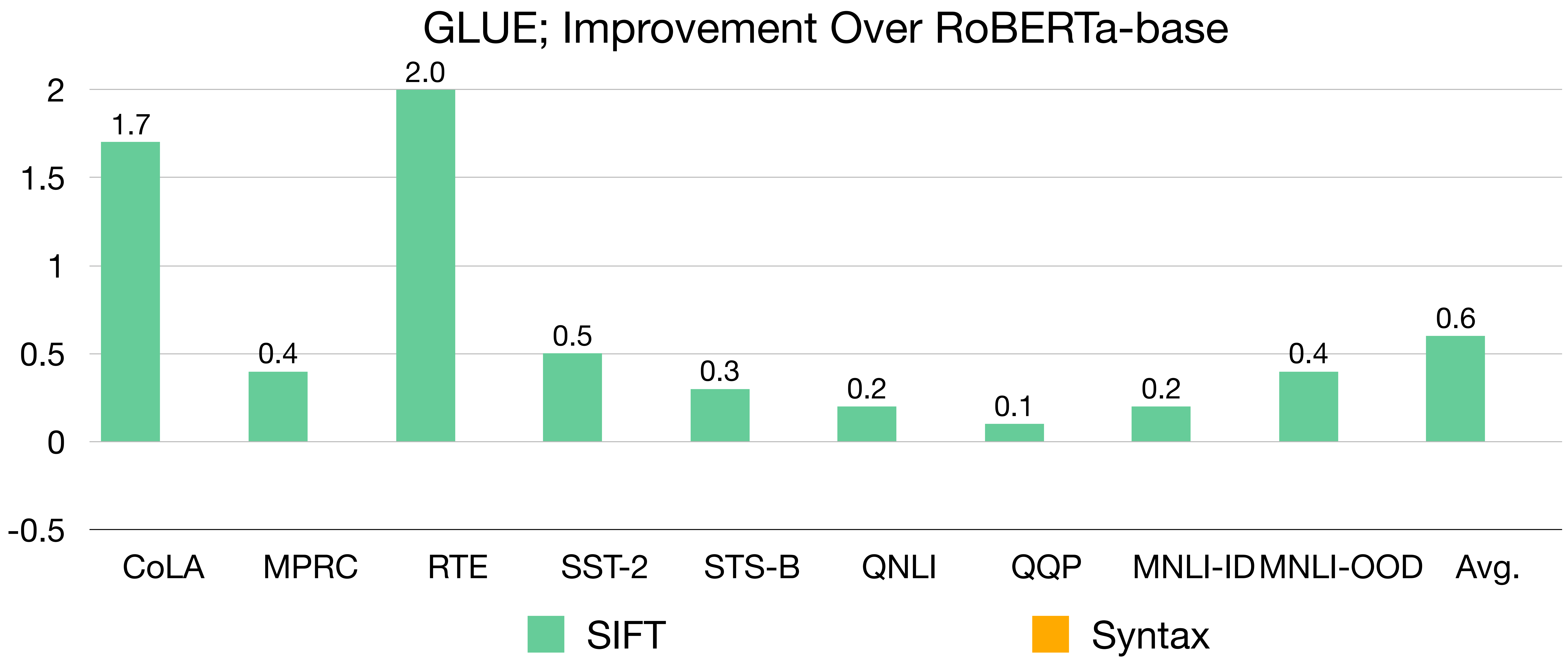


Experiments

- Dataset: GLUE (Wang et al., 2018)
- Backbone: RoBERTa (Liu et al., 2019b)
- Parser: SOTA DM parser with 92.5 labeled F1 (Che et al., 2019)
- Graph Encoder: RGCN (Schlichtkrull et al., 2017)
    - 2 layers
    - Hidden dimension $\in \{256, 512, 768\}$
- Epochs $\in \{3, 10, 20\}$, learning rate $\in \{1 \times 10^{-4}, 2 \times 10^{-5}\}$

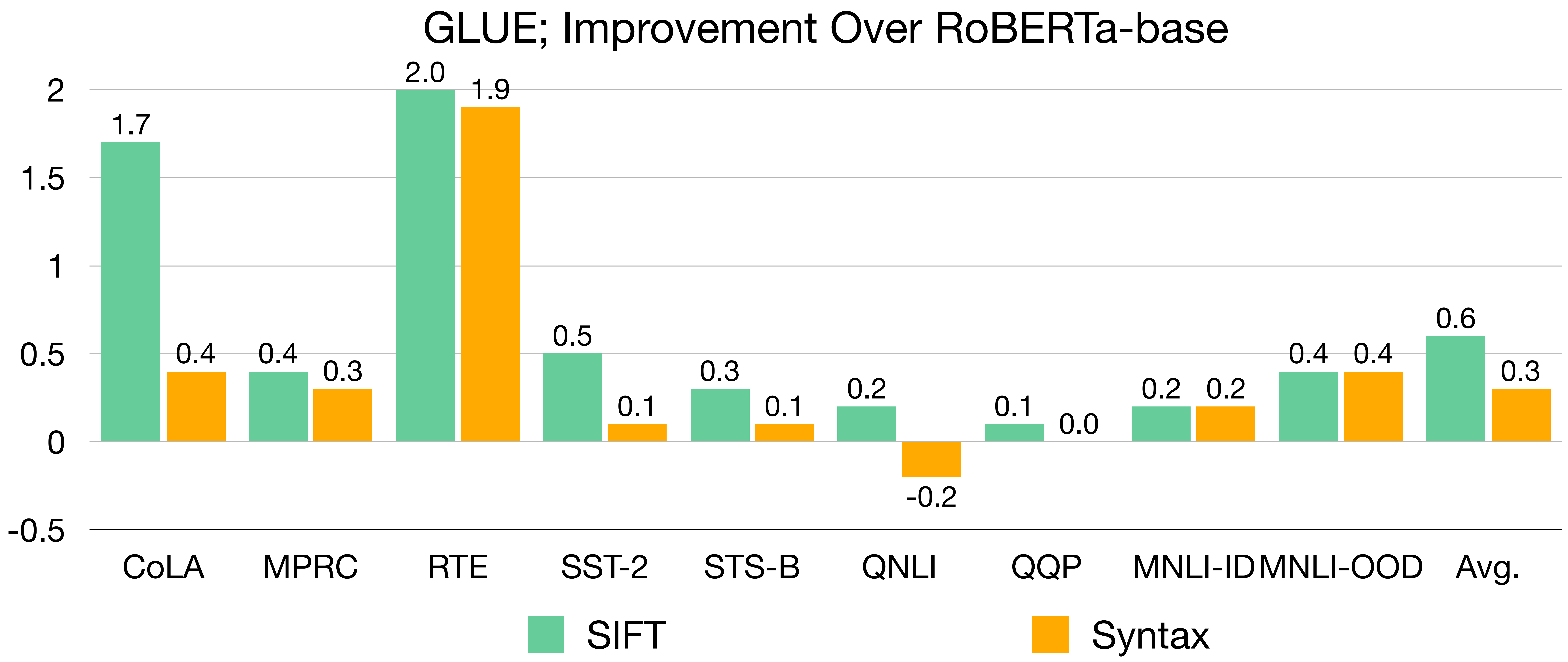
Results

Results



GLUE; Improvement Over RoBERTa-base
2
1.5
1
0.5
0
-0.5
1.7
0.4
2.0
0.5
0.3
0.2
0.1
0.2
0.4
0.6
CoLA
MPRC
RTE
SST-2
STS-B
QNLI
QQP
MNLI-ID
MNLI-OOD
Avg.
SIFT
Syntax

Results


GLUE; Improvement Over RoBERTa-base
2
1.5
1
0.5
0
-0.5
1.7
0.4
0.4
0.3
2.0
1.9
0.5
0.1
0.3
0.1
0.2
-0.2
0.1
0.0
0.2
0.2
0.4
0.4
0.6
0.3
CoLA
MPRC
RTE
SST-2
STS-B
QNLI
QQP
MNLI-ID
MNLI-OOD
Avg.
SIFT
Syntax

Analysis: When Do Semantic Structures Help?

- Two datasets
  - HANS tests if a model uses **invalid reasoning heuristics** (McCoy et al., 2019)
  - GLUE diagnostics tests the model capability in various **linguistic phenomena** (Wang et al., 2018)
- Examine a model trained on existing NLI datasets with synthetic NLI examples

Analysis: HANS Lexical Overlap

The actor stopped the banker. does not entail **The banker stopped the actor.**

RoBERTa	SIFT
68.1	**71.0** (+2.9)

Analysis: HANS Subsequence

The judges heard the actor resigned. does not entail **The judges heard the actor.**

RoBERTa	SIFT
25.8	**29.5** (+3.7)

Analysis: HANS Constituent

If the actor slept, the senator ran. does not entail **The actor slept.**

RoBERTa	SIFT
37.9	37.6 (-0.3)

Analysis: HANS Constituent

If the actor slept, the senator ran. does not entail **The actor slept.**

Before the actor slept, the senator ran. entails

RoBERTa	SIFT
37.9	37.6 (-0.3)

Analysis: GLUE Diagnostics

	Premise		Hypothesis	RoBERTa	SIFT
Pred-Arg Structure	**I opened the door.**	entails	**The door opened.**	43.5	**44.6** (+1.1)
		does not entail	**I opened.**		
Logic	**I have no pet puppy.**	entails	**I have no corgi pet puppy.**	36.2	**38.3** (+2.1)
		does not entail	**I have no pet.**		

Analysis: GLUE Diagnostics

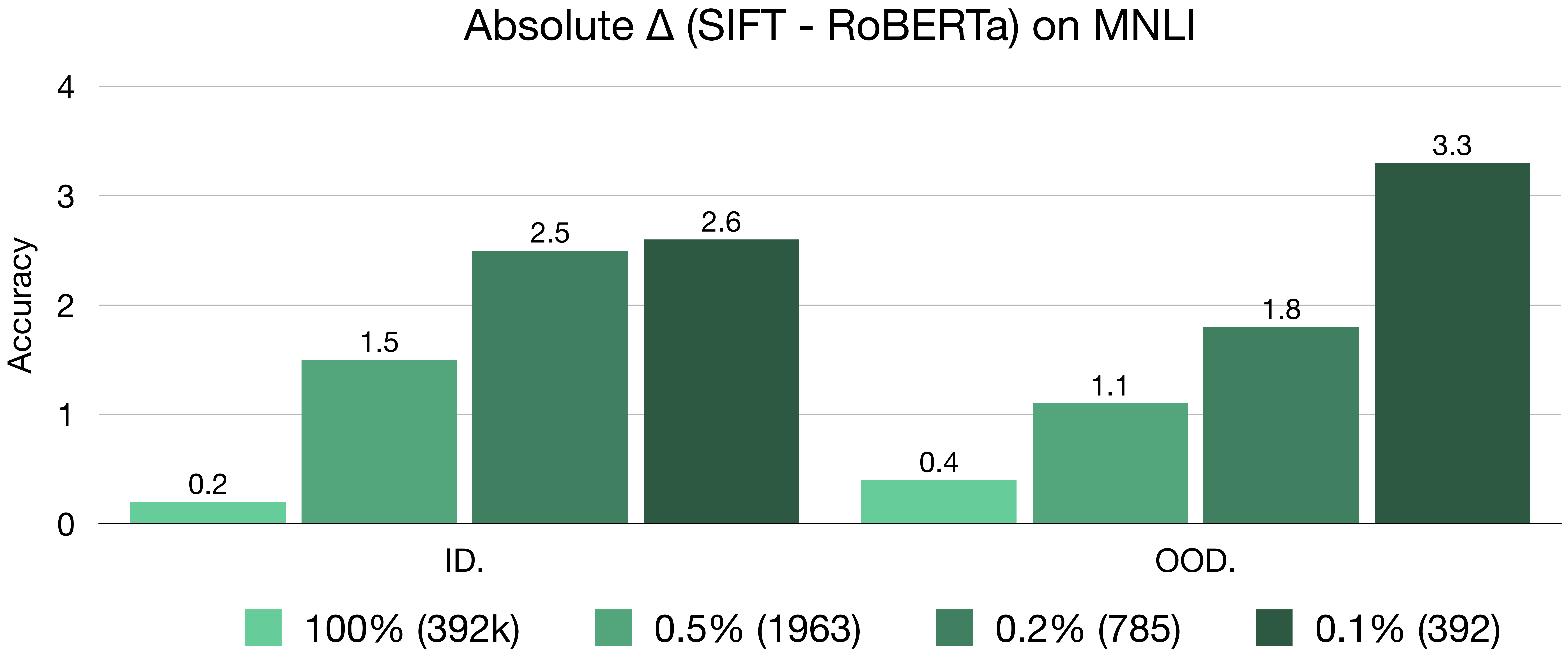
				RoBERTa	SIFT
Lexical Semantics	I have a dog.	entails	I have an animal.	**45.6**	44.8 (-0.8)
		does not entail	I have a cat.		
Knowledge	I live in Seattle.	entails	I live in the U.S.	**28.0**	26.3 (-1.7)
		does not entail	I live in Antarctica.		

Analysis: Sample Efficiency

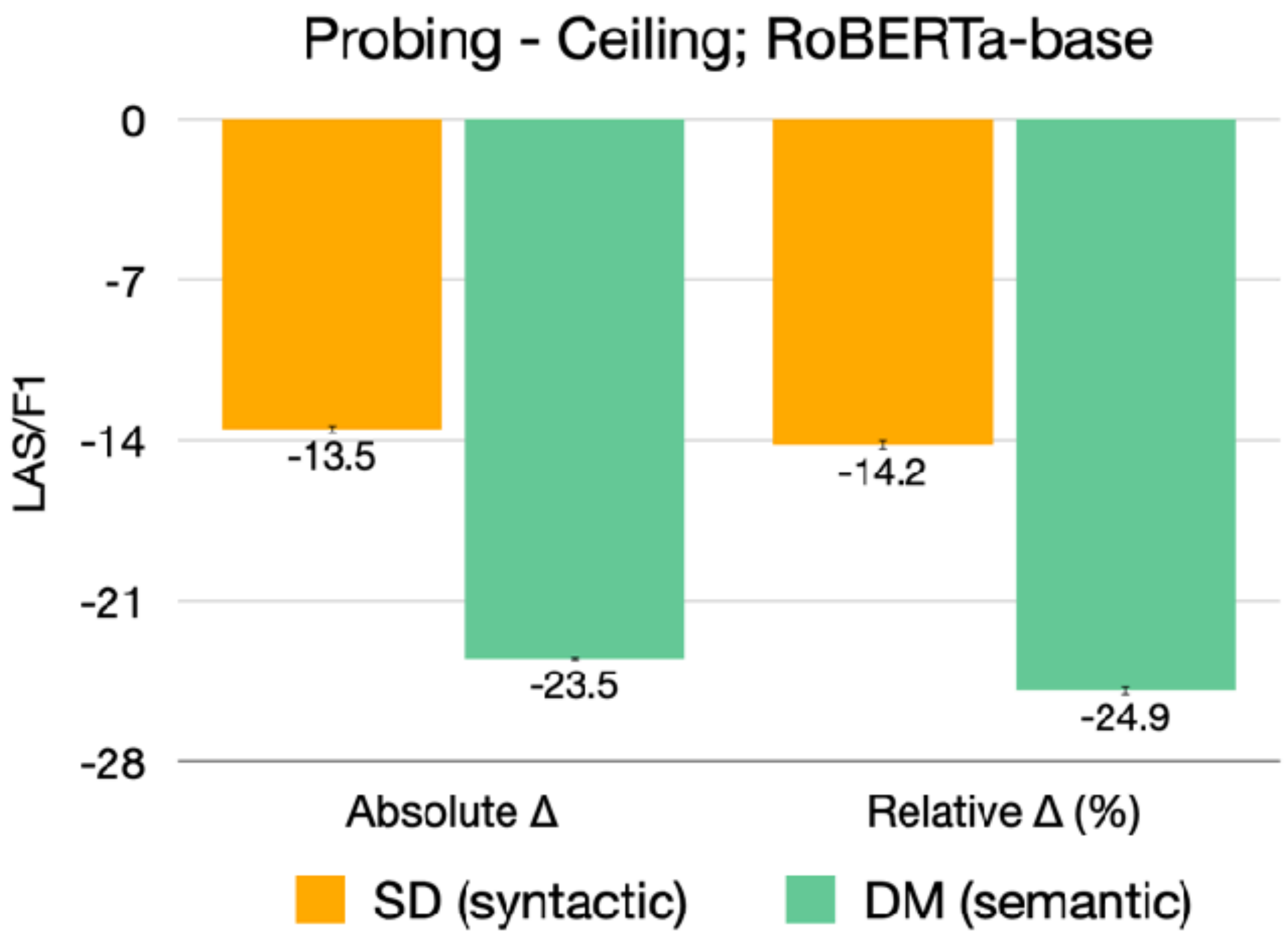
- Use the same downsampled MNLI training set to train RoBERTa & SIFT

Analysis: Sample Efficiency

- Use the same downsampled MNLI training set to train RoBERTa & SIFT

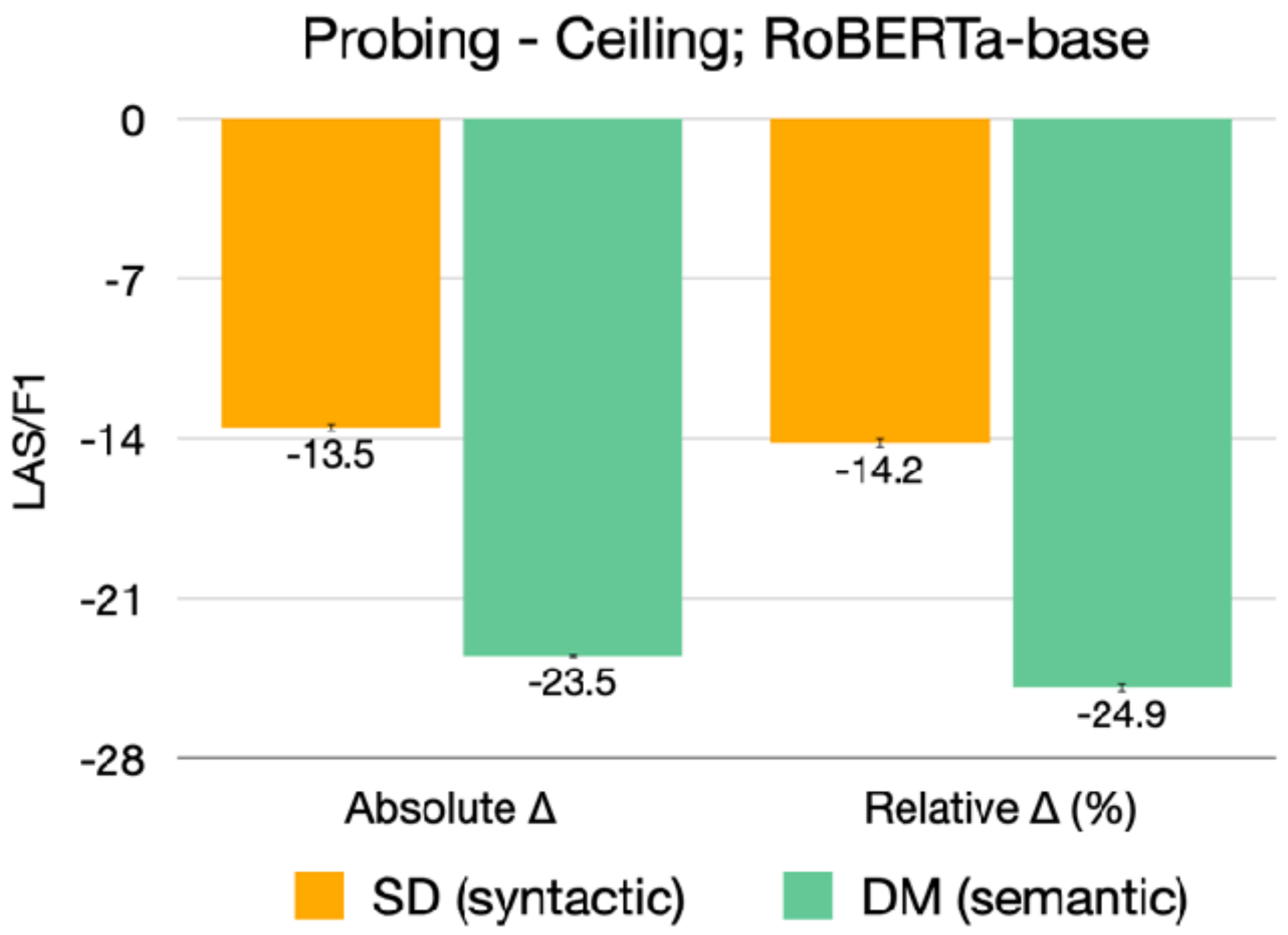


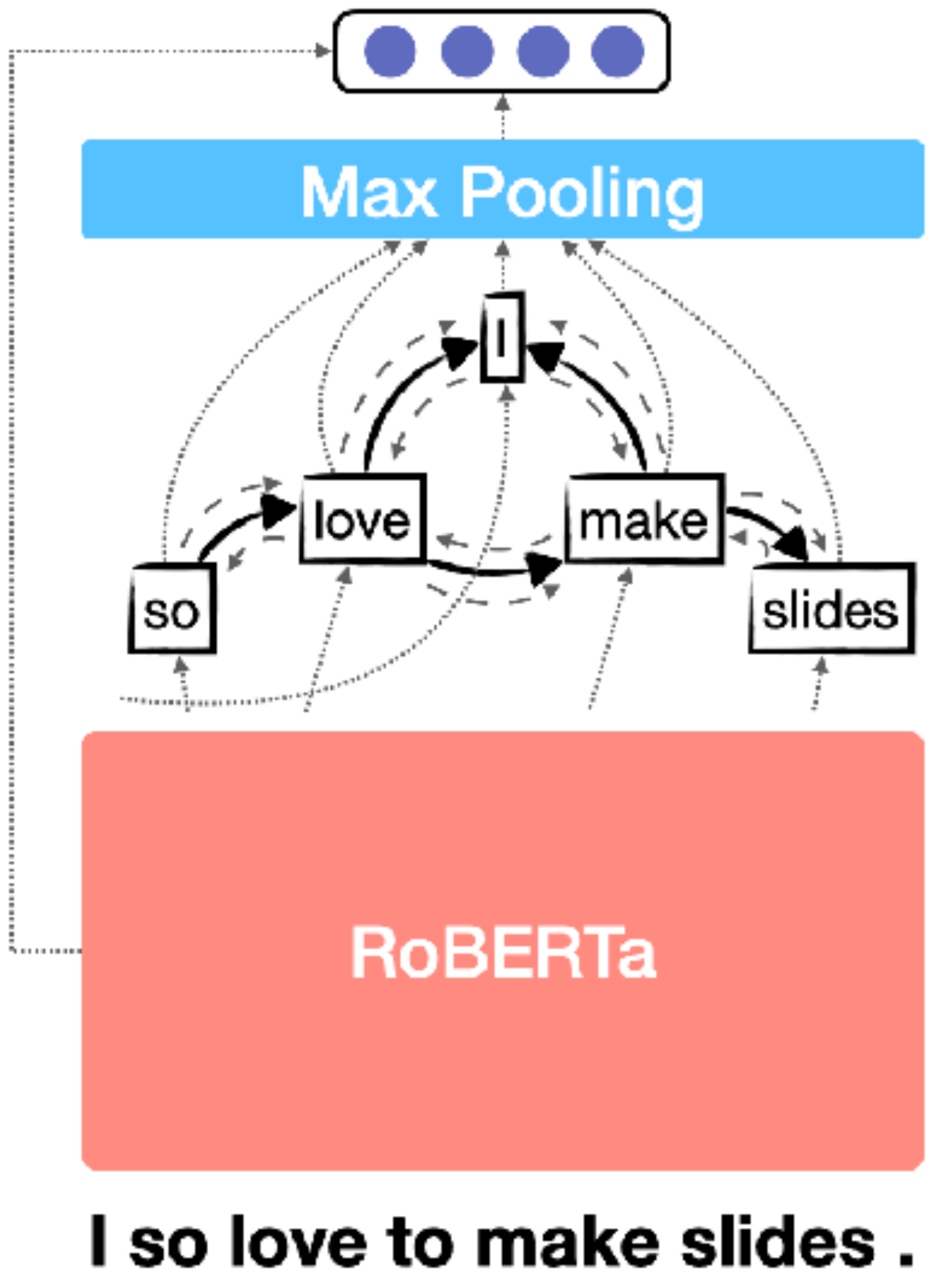
Summary



Probing - Ceiling; RoBERTa-base
LAS/F1
0
-7
-14
-21
-28
-13.5
-23.5
-14.2
-24.9
Absolute Δ
Relative Δ (%)
SD (syntactic)
DM (semantic)

Summary


Probing - Ceiling; RoBERTa-base
0
-7
-14
-21
-28
LAS/F1
-13.5
-23.5
-14.2
-24.9
Absolute Δ
Relative Δ (%)
SD (syntactic)
DM (semantic)


Max Pooling
I
love
make
so
slides
RoBERTa
I so love to make slides .

Thank you!

arxiv.org/abs/2012.05395

github.com/ZhaofengWu/SIFT